Translational Bioinformatics 10
Series Editor: Xiangdong Wang, MD, Ph.D.

John J. Hutton *Editor*

Pediatric Biomedical Informatics

Computer Applications in Pediatric Research

Second Edition

Springer

Translational Bioinformatics

Volume 10

Series editor
Xiangdong Wang, MD, Ph.D.
Professor of Medicine, Executive Director of Zhongshan Hospital Institute of Clinical Science, Fudan University Shanghai Medical College, Shanghai, China
Director of Shanghai Institute of Clinical Bioinformatics, (www.fuccb.org)

Aims and Scope

The Book Series in Translational Bioinformatics is a powerful and integrative resource for understanding and translating discoveries and advances of genomic, transcriptomic, proteomic and bioinformatic technologies into the study of human diseases. The Series represents leading global opinions on the translation of bioinformatics sciences into both the clinical setting and descriptions to medical informatics. It presents the critical evidence to further understand the molecular mechanisms underlying organ or cell dysfunctions in human diseases, the results of genomic, transcriptomic, proteomic and bioinformatic studies from human tissues dedicated to the discovery and validation of diagnostic and prognostic disease biomarkers, essential information on the identification and validation of novel drug targets and the application of tissue genomics, transcriptomics, proteomics and bioinformatics in drug efficacy and toxicity in clinical research.

The Book Series in Translational Bioinformatics focuses on outstanding articles/chapters presenting significant recent works in genomic, transcriptomic, proteomic and bioinformatic profiles related to human organ or cell dysfunctions and clinical findings. The Series includes bioinformatics-driven molecular and cellular disease mechanisms, the understanding of human diseases and the improvement of patient prognoses. Additionally, it provides practical and useful study insights into and protocols of design and methodology.

Series Description

Translational bioinformatics is defined as the development of storage-related, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is the newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders including biomedical scientists, clinicians, and patients. Issues related to database management, administration, or policy will be coordinated through the clinical research informatics domain. Analytic, storage-related, and interpretive methods should be used to improve predictions, early diagnostics, severity monitoring, therapeutic effects, and the prognosis of human diseases.

Recently Published and Forthcoming Volumes

Genomics and Proteomics for Clinical Discovery and Development
Editor: György Marko-Varga
Volume 6

Allergy Bioinformatics
Editors: Ailin Tao, Eyal Raz
Volume 8

Computational and Statistical Epigenomics
Editor: Andrew E. Teschendorff
Volume 7

Transcriptomics and Gene Regulation
Editor: Jiaqian Wu
Volume 9

More information about this series at http://www.springer.com/series/11057

John J. Hutton
Editor

Pediatric Biomedical Informatics

Computer Applications in Pediatric Research

Second Edition

Springer

Editor
John J. Hutton (1937–2016)
Children's Hospital Research Foundation
Cincinnati, Ohio, USA

ISSN 2213-2775 ISSN 2213-2783 (electronic)
Translational Bioinformatics
ISBN 978-981-10-1102-3 ISBN 978-981-10-1104-7 (eBook)
DOI 10.1007/978-981-10-1104-7

Library of Congress Control Number: 2016941065

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media Singapore Pte Ltd.

Foreword

Apps for Pediatrics: Using Informatics to Facilitate, Optimize, and Personalize Care

I was born before electronic computers existed. Two key tools of my early educational years were a slide rule and a typewriter. Just as I emerged from clinical training as a pediatric cardiologist and research training in biochemistry in 1977, calculators and word processors became tools to facilitate clinical care, laboratory research, and medical and scientific communication. I was fascinated by Mendelian disorders and wondered why congenital heart defects sometimes ran in families, but DNA sequencing and other tools to approach genetics did not exist. Only 40 years later, medicine, including pediatrics, and biomedical research, especially genetics and genomics, have been totally changed by biomedical informatics, such that even our American president and other international leaders can propose with confidence initiatives to provide personalized and precision medicine to each individual.

Even with the next generation of tools, from DNA and RNA sequencing to natural language processing of patient and physician notes through artificial intelligences, however, the essential questions of how to apply these current tools to optimize learning and patient care remain. In this second edition of *Pediatric Biomedical Informatics*, with the guidance and tutelage of John Hutton as editor and with the expertise of the faculty and their colleagues who have contributed chapters, the "apps" and approaches to optimize learning, assess clinical outcomes on a population scale, aggregate genomic data for huge numbers of individuals, and organize all of the "big data" created by patients in electronic medical records are explored and presented. The topics discussed include the major core informatics resources needed, from the EMR itself to transmission of information in it for storage and management to security required for patient protection, creation of usable patient data warehouses, and integration of patient information (the phenotype) with biobanked tissue and DNA for research, all critical infrastructure to optimize care and research. In addition, some intriguing "apps" in both patient-oriented research and basic science are provided, to illustrate how population-based studies, assessment

of language, support for decisions, and generation of networks can be done. These "apps" focus on perinatal, neonatal, and pediatric needs. An emphasis on larger, multi-institutional networks and distributed research networks is apparent and essential if we are to more quickly assess the genomics, epigenomics, environmental impact, and treatment outcomes of the relatively rare disorders that we see in pediatrics.

Biomedical informatics is the key to our future, as we integrate clinical care, genomics, and basic science to improve outcomes and discover new therapeutics. With careful design and acquisition of information, we can tame the avalanche of data. We can link and integrate data across institutions to achieve greater power of analysis and increase the speed of discovery and evaluation of treatments. This book provides insight into how to use data to benefit children around our world through "apps" for pediatrics.

Arnold W. Strauss

Department of Pediatrics
University of Cincinnati College of Medicine
Cincinnati, OH, USA

Cincinnati Children's Research Foundation
Cincinnati Children's Hospital Medical Center
Cincinnati, OH, USA

In Memoriam

John J. Hutton, MD, a pioneer and visionary leader across the gamut of Biomedical Research for nearly 50 years, died on June 19, 2016, after a brief but frightening and rapidly progressive form of Amyotrophic Lateral Sclerosis that began to appear over his last year.

Dr. Hutton brought enthusiasm, energy, and effectiveness to virtually all his endeavors, and as some of his physical means were becoming difficult, his energy was particularly fierce about finishing this very book. For this edition, he was passionate that it should address critical issues in biomedical informatics to improve data collection, integration, analysis, discovery, and translation.

Dr. Hutton's career led him to be both a scientific and administrative leader within and above many groups that had specialization within specific areas across the entire process of biomedical research, clinical care, education, and even how to balance the costs of doing all this. He saw the potential of informatics as a natural means of advancing medicine and human health and embraced its mission to build tools, collect and distill data and observations, and to fruitfully carry out, collaborate with, or enable others to perform analyses and propagate significant data and knowledge. That achieving these missions could provide resources to entire communities of educators, researchers, practitioners, and the public raised its significance in Dr. Hutton's view. And he realized this when he graciously asked me if he could come back to do a postdoctoral research project in my computational biology group back in 2003, just after he stepped down after serving as University of Cincinnati Medical School Dean for 15 years. After getting a couple of research projects done and published that mapped and analyzed the significance of gene expression and gene regulatory regions associated with immune cells, tissues, and disease states, and taking a few classes in programming, he was ready to take on running my department! And then from 2005 to 2015 he served as Bioinformatics Division Director and Senior Vice President for Information Technology at Children's Hospital Medical Center, the oversized Pediatrics Department for the College of Medicine.

A native of eastern Kentucky, Dr. Hutton graduated in Physics from Harvard and attended the Rockefeller University and Harvard Medical School where he obtained

an MD degree and completed postgraduate training in internal medicine with a research focus in biochemistry, genetics at the National Heart Lung and Blood Institute, and clinical training in hematology-oncology at the Massachusetts General Hospital. From 1968 to 1971 he served as a Section Chief at the Roche Institute for Molecular Biology, then 1971–1980 as Professor and Medical Service Chief at the University of Kentucky, then 1980–1984 as Professor and Associate Chief at the University of Texas Medical Center San Antonio, and from 1983 to 1988 as a member and Chair of the famous NIH Biochemistry Study Section. Dr. Hutton returned closer to his original Kentucky home in 1984 as the Albert B. Sabin Professor and Vice Chairman, Department of Pediatrics, Cincinnati Children's Hospital Medical Center. In 1987, he was appointed Dean of the College of Medicine, and he served in that role up until 2003, also serving national roles in the American Society of Hematology, and as Executive Council Member for the Association of American Medical Colleges. All of this experience contributed to his understanding the nature of the multidisciplinary problems that computational biology could, and should, solve.

Dr. Hutton's research career included editorial oversight of textbooks in internal medicine and pediatric bioinformatics and more than 200 peer-reviewed papers, including among the first trials of gene therapy for inborn immunodeficiency.

As Dean, he was principal investigator of a multimillion dollar Howard Hughes infrastructure improvement grant, which focused on development of resources in genomics, proteomics, and bioinformatics. And after stepping down as Dean, he won a $1.7 million IAIMS grant from NIH/National Library of Medicine, which was awarded to support innovative research in and development of information management systems. Other of Dr. Hutton's passions included the College of Medicine's MD-PhD Physician Scientist Training Program, which was nationally recognized for its high quality and received peer-reviewed funding from the NIH, and also a strong emphasis on the development of programs in Ethics for Medicine and Medical Research. An endowed annual Hutton Lectureship was established in Medical Ethics, and an endowed Hutton Chair in Biomedical Informatics was established for Cincinnati Children's Hospital Research Foundation, and I am extremely honored to be its first recipient.

Dr. Hutton's family includes his wife, Mary Ellyn, a classical musician who also writes about classical music for the Cincinnati Post and other publications. His daughter, Becky, graduated from the UC College of Nursing, married Thomas Fink, has four children, and lives in Tipp City, Ohio. His son, John, graduated from Davidson College and the UC College of Medicine, married Sandra Gross, has two children, and lives in Mt. Adams. His daughter, Elizabeth, graduated from Harvard in 2001, works in Boston, and will enter the Ohio State University College of Law in August 2003.

Dr. Hutton leaves us all a rich legacy of achievement and inspiration to be and empower the next generation of computationally empowered students, researchers, educators, and practitioners.

Bruce Aronow, PhD, the John J Hutton, MD Professor of Biomedical Informatics, University of Cincinnati Department of Pediatrics, Department of Biomedical Informatics, Cincinnati Children's Hospital Medical Center.

Contents

Part I
Core Informatics Resources

Chapter 1
Electronic Health Records in Pediatrics

S. Andrew Spooner and Eric S. Kirkendall

Abstract Most pediatric healthcare providers use an electronic health record (EHR) system in both office-based and hospital-based practice in the United States. While some pediatric-specific EHR systems exist for the office-based market, the majority of EHR systems used in the care of children are designed for general use across all specialties. Pediatric providers have succeeded in influencing the development of these systems to serve the special needs of child health (e.g., immunization management, dosing by body weight, growth monitoring, developmental assessment), but the pediatric community continues to press for further refinement of these systems to meet the advanced needs of pediatric specialties. These clinical systems are typically integrated with administrative (scheduling, billing, registration, etc.) systems, and the output of both types of systems are often used in research. A large portion of the data from the clinical side remains in free-text form, which raises challenges to the use of these data in research. In this chapter, we discuss workflows with data implications of special importance in pediatrics. We will also summarize efforts to create standard quality measures and the rise in EHR-based registry systems.

Keywords Pediatric EHRs • EHR market • Pediatic workflow • Growth charts • Pediatric drug dosing • Developmental monitoring • Immunizations

---

S.A. Spooner, M.D., M.S., FAAP (✉)
Departments of Pediatrics and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine,
3333 Burnet Avenue, MLC-9009, Cincinnati, OH 45229, USA
e-mail: andrew.spooner@cchmc.org

E.S. Kirkendall, MD
Departments of Pediatrics and Biomedical Informatics, Divisions of Hospital Medicine and Biomedical Informatics, Cincinnati Children's Hospital Medical Center,
University of Cincinnati College of Medicine,
3333 Burnet Avenue, MLC-3024, Cincinnati, OH 45229, USA

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_1

1.1 Current State

To understand the inherent challenges and potential of using pediatric EHRs for research, one must understand the extent to which EHR systems are used in various pediatric settings. Information on adoption of these systems in child health is best known for the United States, but trends are similar in other developed countries.

1.1.1 Adoption Rates

Child health providers—specifically, pediatricians—are thought to be slower than general practice providers in adopting electronic health record technology (Lehmann et al. 2015; Leu et al. 2012; Nakamura et al. 2010). Children's hospitals, because they tend to be urban and academic, are often ahead in adoption as are institutions of larger size (Andrews et al. 2014; Nakamura et al. 2010). The reason for the slower adoption in pediatric practices probably relates to the difficulty of fitting child health needs into a product designed for general, adult care. In this way, current EHRs violate the pediatric adage that "children are not small adults." If EHRs are not designed or cannot accommodate the unique needs of the pediatric population, healthcare providers are not likely to be quick adopters of such systems. A recent estimate of pediatric adoption of fully-functional EHRs in ambulatory practice are at about 6 % (Leu et al. 2012), although by now this is undoubtedly higher given recent trends (Lehmann et al. 2015).

The U.S. Meaningful Use program of the HITECH Act (Blumenthal and Tavenner 2010; HHS 2009) of the American Recovery and Reinvestment Act of 2009 intends to provide financial stimulus to physicians and hospitals to adopt EHR technology. There is a version of the program for Medicare (the the U.S. federal public payment system for the elderly) and for Medicaid (the U.S. state/federal program for the poor and disabled). Since pediatricians do not generally see Medicare patients, child health providers and hospitals usually qualify for the Medicaid program. In this program, individual providers may qualify for an incentive payment if they have a minimum of 30 % Medicaid patient volume, or, if they are pediatricians, 20 % Medicaid volume. This criterion covers about half of the office-based pediatricians in the United States (Finnegan et al. 2009) but does leave out a significant number with very low Medicaid volumes. These providers tend to practice in more affluent areas, but pediatrics is not a specialty with very high margins under the best of circumstances, so Meaningful Use will not directly affect the adoption rates for this large group. Member survey data from the American Academy of Pediatrics estimate that up to 2/3 of U.S. pediatricians may be eligible for some incentive payment (Kressly 2009), so the next few years may be a time of rapidly increasing pediatric deployment of EHRs.

1.1.2 The Pediatric EHR Market

The pediatric EHR market includes small pediatric practices of one or two practitioners all the way up to large, hospital-owned practices of hundreds of pediatricians. There is similar variability in the crowded U.S. EHR vendor market, where a given company specializes in offering its product to practices of a certain type or specialty area. In the early 1990s, almost all electronic medical record systems were of the home-grown variety. Today, several hundred companies in the U.S. market offer over 3000 different EHR systems (ONC 2016) and the services that accompany their deployment, customization, and maintenance. While there has been some vendor dropout and consolidation (Green 2015), the EHR marketplace is far from the point where only a few major companies service the majority of customers. Because of the small size of the pediatric EHR market, there have been very few companies that have succeeded in marketing a product that is specific to pediatrics.

1.1.3 Vendor Systems

EHR systems today are sold by software vendors attempting to gain a large enough customer base to sustain a business. While this model provides a more sustainable source for software than the home-grown model, it creates a problem for child health providers: Most customers are not pediatric, so most EHRs are not designed specifically for pediatric care. A further problem for child health researchers is that practically none of these systems are designed with research in mind. Instead, they are designed for patient care and the administrative tasks that support patient care. Figure 1.1 is a mock-up of an EHR screen that highlights these assumptions.

While these assumptions are not truly prohibitive of these systems' use in a pediatric environment, they often force workarounds that affect the quality of data in the system. For example, when faced with the unavailability of an adequate field to capture a concept, one may feel forced to use a free-text field intended for some other purpose to store the information. In this case the data loses integrity (such as a conversion from structured to unstructured data) and it becomes impossible to apply computational methods to the data.

Child health professional groups have attempted to promulgate catalogs of functionality necessary for the care of infants and children (AHRQ 2013; CCHIT 2010; Kim and Lehmann 2008; Spooner 2007; Spooner and Classen 2009). Fortunately, vendors who sell systems to pediatric practices and children's hospitals are gradually creating mature systems that respond to their customers' pediatric-specific needs.

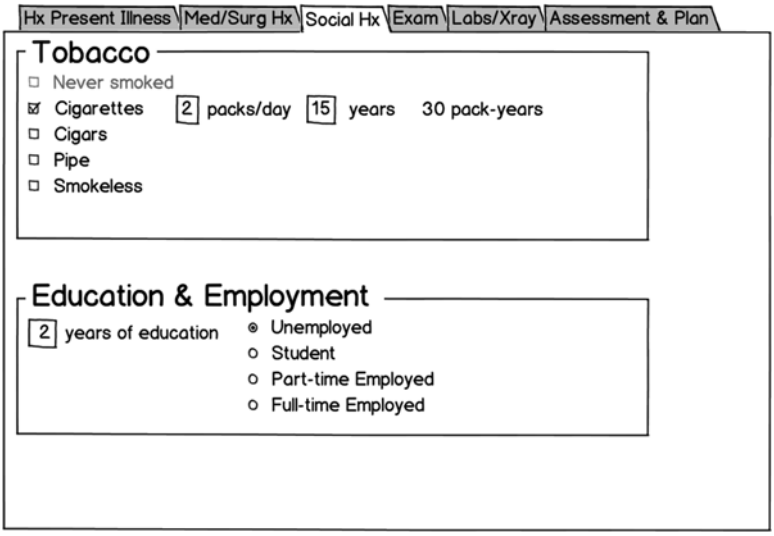


Fig. 1.1 Elements of an EHR user interface that imply an exclusive orientation to adult patients. In the case of tobacco history for an infant, one would be interested in recording passive smoke exposure, which is not included in this display. In the education section, it is implied that one's years of education are fixed and in the past, as they would be for most adults

1.1.4 Homegrown Systems and Publication Bias

Despite the prevalence of vendor systems in the marketplace, the bulk of reported literature on the use of EHRs from the initial reports of the 1970s through most of the first decade of the 2000s is based on experience with home-grown systems (Friedlin et al. 2007; Gardner et al. 1999; Miller et al. 2005). The result is that the evidence on which to guide the implementation of EHRs is only partially applicable to most installed systems. Add to this the complexities of systems customized for pediatric care and the connection between the results of the adult-oriented, home-grown software and installed, vendor-provided systems is even more tenuous. This phenomenon makes the pediatric EHR environment ripe for research to be conducted on the systems themselves, but it also makes it hard to definitively answer questions about what works best. As such, reports in the informatics literature should be critically analyzed to determine the external validity of published results, in particular whether the system being tested or described is a vendor solution or homegrown application.

1.1.5 Pediatric Versus General Environments

The main features that differentiate the pediatric environment from that of general adult care are:

- ***Diseases And Conditions That Are More Prevalent In The Young;*** Congenital disease and conditions related to abnormal growth and development are not usually part of adult care. Templates, data fields, terminology systems, and other clinical content in an EHR may therefore require customization to meet different clinical needs.
- ***Parental/Guardian Proxy;*** In the pediatric environment parents (or guardians) are almost always involved in encounters and responsible for care decisions. While there are certainly family members of adults involved in the care of the patient, in most cases the patient is competent to make health care decisions. Siblings may receive care at the same encounter.
- ***Physical And Developmental Growth;*** The pediatric patient is growing and developing physically and mentally at a fast clip. Weights change rapidly, especially in the first year of life. Developmental capability to participate in self-care increases with age. Because of children's dependent status, social situation has a much greater impact on health than in most adult care.

1.1.6 Pediatric Subspecialties Versus the General Purpose EHR

If it were not difficult enough to apply pediatric assumptions to general-purpose systems, the difficulty is compounded in the case of pediatric specialties. Specialty care entails more detailed, less common, and often more granular, special requirements. There is also more variation of care practices at the subspecialty level as there tends to be less evidence available to standardize procedures and protocols. It is not uncommon for several physicians within the same group to have differing opinions on best practices when little evidence exists to guide the way. In many cases there may also be a paucity of pediatric research (as compared to adults), further complicating the issues of standardization.

In pediatric specialties, the very clinical content of the practice may be quite different from their adult counterparts. Pediatric cardiology, for example, is chiefly concerned with congenital disease, whereas adult cardiology focuses more on acquired cardiovascular disease. This shifting of focus on disease etiology and pathology disallows any loose extrapolation and adoption of adult data to the pediatric population.

1.1.7 Data from Natural Workflow vs. Research, Primary vs. Secondary Use of Data

As EHRs are designed to support clinical care, data that makes its way from the EHR into a data repository is of lower quality than what one might find in data specifically collected for research. Data validation, completeness, and standard

processes are very much secondary to successful completion of clinical work. It is for this reason that most research from clinical environments is based on claims data, where some energy is expended to ensure data accuracy and completeness. Of course, claims data is at least one step removed from the important clinical details of the patient encounter.

1.2 Workflow and the EHR

1.2.1 Data Entry

The function of an EHR is not primarily to serve as a data-entry tool. Its purpose is to facilitate patient care for individual patients. In doing so it offers some opportunities for data extraction for other purposes (operations analysis, research, quality measurement). Since EHRs are not designed for research, analytics, or population management, there will always be a need to input research-specific EHR data into the data repository, as well as methods to extract it. The value of that data is directly related to the quality of the data entry. Missing values threaten the validity of any measures based on the data and data cleansing, a time and resource-consuming endeavor. For this reason, it is best to use data that is already captured reliably (like orders for tests) or to make workflow changes to increase reliable data entry. In a busy clinical environment where clinicians are already working at capacity to meet documentation guidelines for billing, there is little opportunity to make these changes. Clinicians will often ask for a "hard stop" reminder to enter data (or, more commonly, to get someone else to enter data), but the effectiveness of alerts is very limited (Strom et al. 2010) and hard stops are usually abandoned as annoying. Any effort to make sense of the quality and integrity of EHR data must take into account some knowledge of the clinical workflows that produced it.

1.2.2 Multiple Job Roles and Their Interaction with the Record

Like the paper record, the electronic record accepts input from people in multiple job roles: physician, advanced-practice nurse, physician assistant, nurse, medical assistant, and multiple clerical roles, among others. Effective data organization depends on clear job roles related to the record. For example, if it is not clear who is responsible for the accuracy of the patient's medication list, the data extracted will be of low quality. When one puts together a plan for the use of EHR data, part of the workflow analysis should include the establishment of how clear the job roles are. Job roles, or "profiles" as EHR systems refer to them, usually define how data is viewed and input in the user interface. When this variation occurs, it is not unusual for data to be entered (or not entered) in multiple ways. Great attention should be

paid in designing or customizing these screens and standardization of entry and viewing carried out whenever possible.

1.2.3 Special Pediatric Workflow Issues

Multiple Patients Per Encounter Siblings within the same family are often seen together, especially for well-child care. In no other area of medicine is this type of multi-encounter a common experience. EHRs can be configured to allow access to multiple patient records at once, but data sharing between patients is not typically supported. In the pediatrics, there are areas of EHR data that ought to be shared between patients, like family history and social history, or guarantor information, but typically this must be entered separately for each patient.

One example where linking of records could be helpful in both the adult and pediatric EHR would be to provide the capability to update and/or duplicate the family history section in related patients' charts. For instance, if two siblings are taken care of by the same practice, family history in their direct ancestors would be identical. If the records were linked through an EHR social network, updated data in one sibling's chart could offer a prompt in the other sibling's chart that useful information needs to be verified and inserted into the record. This form of networking could also prove helpful in generation of genograms. In a more practical fashion, duplication of pregnancy circumstances and perinatal events in the charts of twins would reduce large amounts of manual data entry. There are a variety of medico-legal and ethical concerns with these kind of linkages that will not be addressed here, but the reader should be aware of the current paucity of this functionality and its implications in research data obtained from EHRs.

Multidisciplinary Clinics The large number of rare disorders seen in pediatrics, coupled with the relative rarity of pediatric specialists with expertise in these disorders, creates the need to bring multiple specialists together into a single patient encounter. Arranging visits this way is a great convenience to the family, but also allows specialists to work together on difficult multi-organ problems that might otherwise take months to coordinate. In children's hospitals, numerous clinics of these type are created or the constituents modified every year. EHR systems should support this kind of workflow, but since it is not typical in adult or non-specialty care, it is not a smoothly implemented, standard feature of most EHRs.

1.3 Special Functional Requirements and Associated Data

The following section describes some of the functionality and associated data requirements that are, for the most part, unique to pEHRs. We discuss both basic functionality that should be considered required, as well as optimal, ideal functionality that would greatly increase the data quality captured in EHRs.

1.3.1 Growth Monitoring (Including Functions of Interest Only to Specialty Care); Basic Growth-Chart Functionality

Perhaps the one clinical activity that distinguishes pediatric care from adult care is growth monitoring. While weights, skinfold measurements, and even height are measured in adult care and tracked, the assumption is that these are stable measures. Growth and development are fundamental processes in pediatrics, especially in the ambulatory setting. The rapid progression in both are carefully tracked in the longitudinal health records and constantly evaluated for normality or deviation from expected patterns. As such, it is expected that optimal EHRs have the ability to robustly track and identify both healthy and pathologic growth. In children, of course, there are growth patterns that constitute a wide range of normal, and growth patterns that signify disease. Some diseases, like growth hormone deficiency, affect growth directly. Others, such as inflammatory bowel disease, affect growth negatively through catabolic and energy-consuming inflammatory processes. Other abnormal growth patterns are part of inherited conditions like Prader-Willi syndrome (obesity) or Turner syndrome (short stature). In routine, well-child care, examination of the growth chart is standard practice. In the ongoing management of specific, growth-affecting conditions, growth chart analysis is similarly routine. EHRs that intend to support pediatric care must support display of these data in a way that goes beyond a simple time plot of the values. Critical to the functioning of a growth chart display is concomitant display of growth norms, in order to allow interpretation of patterns (Rosenbloom et al. 2006).

1.3.2 Data Found in Growth Chart

Weight and **stature** are the very basic data tracked in growth charts, but the concept of height for patients who cannot stand (or stand cooperatively) is usually conceptualized as length. Norms for young children (less than 2 years old) are typically separated from those of older children in this respect. In a typical EHR, there are growth charts for children 0-36 months old and for those over 2 years old. Data storage for the points that are plotted on the stature curves may therefore vary as to which is a height and which is a length. Growth percentiles of the same data point will also vary across different chart types, which can be particularly confusing in the 24–36 month age range. The same height or weight, for example, will often generate discrepant percentiles when a user alternates between views of different growth charts.

See Fig. 1.2 for examples of typical growth charts in use in an EHR. The essential function of the growth chart is to give the user a sense for where the patient falls within a similar age population, expressed as the percentile at that age. Values higher than 95 % or so or below 5 % or so are considered abnormal, but must of

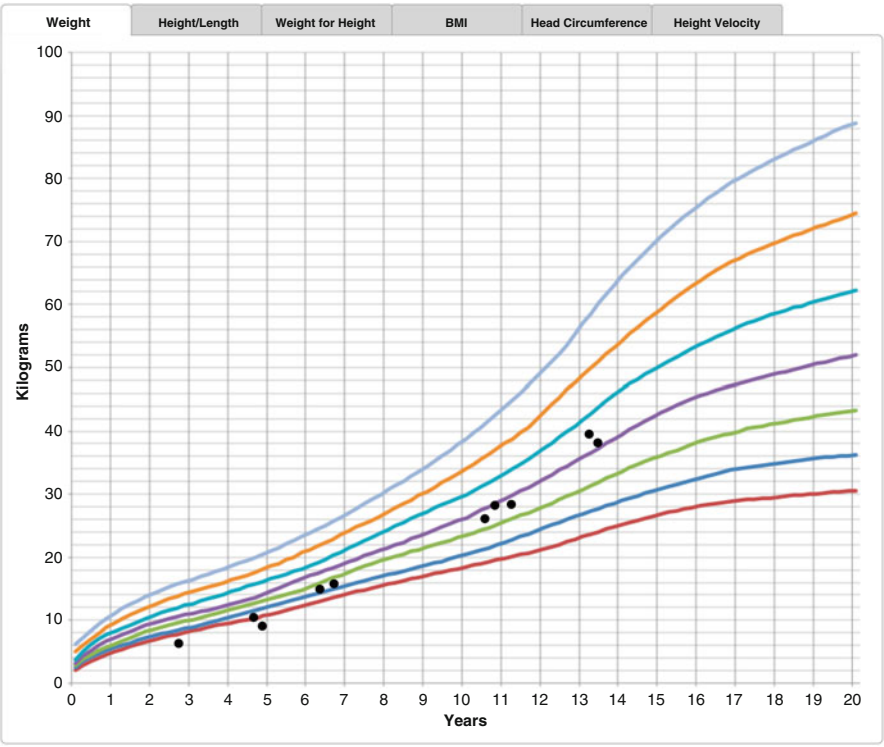


Fig. 1.2 Mockup of a growth chart as deployed in an electronic health record system. The isobars represent constant age-specific percentile for the metric (in this case, weight). In this case the patient has crossed the 3rd, 10th, and 25th percentile isobars. This might represent an abnormal growth pattern (gaining too much weight) or recovery from chronic illness to a normal weight, depending on the clinical situation

course be interpreted in the context of the child's overall growth. For example, if a normal child happens to be small, owing to their genetic predisposition, they may never rise to a particular predetermined percentile. Their growth velocity may be considered normal as it hovers around the 2nd percentile throughout life. Such tendencies are referred to as "following their own curve"; in fact, departures from that curve into "normal" range may indicate an abnormal state for that patient. It is this complexity that makes growth charts irreplaceable by number-driven decision support. There does not appear to be a current substitute for a clinician viewing the curve graphically against a set of norms.

Head Circumference is also essential for basic growth chart functionality. In standard growth charts used in general pediatric care, these charts go up through age 36 months. There are norms for older children and young adults (Nellhaus 1968; Rollins et al. 2010), but these are used only in specialty practices like oncology or neurosurgery to monitor head growth after radiation or tumor removal.

Body Mass Index calculated from weight and stature, is also becoming a standard growth chart in pediatric practice. In adults, when BMI is used as an index of the severity of either obesity or malnutrition, the cutoff values to indicate abnormal body mass index are the same for all ages. In children, interpretation of BMI rests on the percentile value within the child's current age. The U.S. Centers for Disease control publishes these norms (CDC 2012) so that graphs can be made and percentiles calculated.

Height Velocity In specialized areas of pediatrics, where growth is the focus (e.g., endocrinology), there are normative curves, implemented much like the curves for primary anthropometrics, for the rate at which height changes over time. These curves are used to evaluate the severity of growth impairment and to monitor the use of drugs which might affect growth one way or the other. There are no published curves for weight velocity, although the current interest and prevalence of obesity in the U.S. may change that.

Other Anthropometric Values Norms for chest circumference, skinfold thickness, and leg length have been developed, but are used infrequently. In any case, the techniques for display, where data are displayed against normative curves, remain the same.

Percentile/Z-Score Calculations While plotting primary data against norms makes for an effective visual display to support clinical decisions, information systems can compute the applicable percentiles given a measurement and an age, provided the proper normative data are available for the calculation. The U.S. CDC provides tables for this data for the datasets they publish, and a process for computing the percentiles (CDC 2012) (see the WHO vs CDC subsection below). Most growth charts are published merely in graphical form, and the data required to perform the computation is not provided. The computation process calculates a z-score (number of standard deviations above or below the mean for an age) and then applies assumptions about the distribution to come up with a percentile within that distribution. For extremes of growth, the z-score itself may be more useful, since the difference between a weight at the 99.96th percentile may be hard to distinguish from a weight at the 99.99th percentile otherwise. Few EHRs provide the z-score directly, but it is a desired functionality for pediatric specialists who focus on growth.

1.3.3 Special Population Data

Up until now, we have discussed EHR functionality associated with normal growth. In this subsection, we address the topics of collecting and managing special population data.

Congenital Conditions Disordered growth is a major feature of a variety of congenital conditions such as Noonan syndrome (Ranke et al. 1988), Laron dwarfism (Laron et al. 1993), and Williams syndrome (Martin et al. 2007). The measurements are the same, and the growth charts work the same way, but the normative data are different. EHR systems generally provide some of these normative curves that can be selected manually or automatically depending on clinical conditions.

Extremes of Growth In conditions causing extreme failure to thrive or in obesity, the usual normative curves that express curves close to the 99th and 1st percentile may not be adequate. In these cases, the data points are so far removed from the highest and lowest curves that users find it difficult to describe patterns based on the curves. In these cases it is better to create normative curves based on z-scores, so that users can express the patient's growth state relative to the position of these curves far outside the normal range.

Intrauterine Growth Similar to post-natal curves, intrauterine curves, based on gestational age, combined with parameters measurable via ultrasound (crown-rump length for stature or biparietal diameter for head size) are useful for expressing growth patterns of fetuses. These sorts of curves are more often found in system designed for obstetric use, but may be useful in the immediate post-natal age.

WHO vs. CDC The World Health Organization has published a set of growth charts for infants that are based on a sample of healthy, breast-fed infants (Grummer-Strawn et al. 2010) The motivation for creating these charts is to present a more physiologically accurate view of how normal infants should grow. Because the CDC growth data has been in use much longer, EHR system vendors have had to deal with the ambiguity of two widely accepted growth charts for normal infants. Researchers using percentile growth data from EHRs should be aware and take note of the source in order to make accurate comparisons.

Specialized Growth Analysis Growth chart data must sometimes be temporally related to other physiologic events. For example, one may want to indicate on the growth chart a child's sexual maturity rating, since advanced stages of sexual maturation are associated with cessation of normal growth. One might also want to indicate the "bone age" (an estimate of age based on the appearance of bones on plain-film radiography) on the growth chart in cases where the age of the patient is uncertain, as in some cases of international adoption. There are no standard ways of displaying these data within a growth chart, but practitioners who focus on growth cite this function as essential to full growth chart functioning (Rosenbloom et al. 2006).

Correction for Gestational Age Infants born prematurely, because of their smaller size, require special growth charts (Fenton 2003; Fenton and Kim 2013). Outside the immediate post-natal period, though, practitioners generally use regular growth charts, and graphically indicate a correction factor for prematurity. The expectation

is that premature infants will eventually catch up to other infants of the same postnatal age. Part of the analysis of growth charts in premature infants is the time it takes them to achieve this catch-up growth.

1.4 Drug Dosing

Given the inherently changing growth of children, prescribing the appropriate dose of medications can be difficult. What follows is a discussion of the practical and research implications of prescribing medications through an EHR.

Weight-Based Calculations Medications in infants and small children are generally dosed by body weight. As body weight increases with age, children grow into the adult dose; the weight at which one can receive an adult dose varies by medication. Such weight dependence makes the act of prescribing medications to young people more complex. In addition to the act of prescribing, there are complexities related to the storage of the prescription and the decision support that might be provided to the user. EHRs used in the care of children should, at a minimum, provide the ability to calculate a given dose of a medication based on the actual weight (Kirkendall et al. 2014; Shiffman et al. 2000; Spooner 2007). More advanced functionality includes providing standard weight based doses, offering dose range checking, and providing dose ranges dependent on clinical factors, like diagnosis.

Weight Changes As infants grow, their body weight changes rapidly enough that they may "grow out of" a medication at a given dose. Providers who care for infants on chronic medications know to re-prescribe when body weight changes, but a sufficiently sophisticated information system can help to support the decision to re-prescribe, or at least to make it easier by carrying forward weight-based dosages to subsequent prescriptions. Data structures used to store prescriptions must therefore retain the weight-based dose (e.g., 40 mg/kg/day) as data.

Dosing Weight It is not always the case that actual body weight is the best datum to use in calculating weight-based prescriptions. In very young neonates, who lose significant amounts of weight as they adjust to life outside the womb in the postnatal period, one may prefer to use a "dosing weight" to keep prescriptions consistent. Similarly, patients who gain weight due to edema will need a more appropriate weight upon which to base dosing decisions. EHR systems need to take into account different methods of storing and using weight in support of prescribing.

Daily vs. Single-Dose Reasoning In dose-range decision support, there are limits for single doses and for total daily doses, both of which must be accounted for in decision support. Pediatric prescribing guidelines are usually written in mg/kg/day divided into a certain number of doses. This format takes into account the per-dose and daily dose parameters, although EHR dosing rules may provide these two parameters separately.

Power-of-Ten Errors In providing small doses to small people one of the most common and most dangerous dosing errors is the situation where the dose is higher by a factor of 10 or 100, due to confusion between the volume to administer (e.g. 2 mL) and the mass to be administered (20 mg), faulty multiplication, or the migration of a decimal point (Dart and Rumack 2012). In adult care, doses tend to be standard, so there is no way for practitioners to recognize the "usual" dose, since there is no usual dose. Dosing decision support in EHRs is mainly intended to mitigate these kinds of errors.

Physiologic Variation with Development A subtle factor that affects some pediatric prescribing is the effect of maturation of organ systems in the immediate postnatal period that affect drug clearance rates. In order to provide adequate decision support for these dose variations, the ideal system would need to be able to compute different ideal doses at ages measured in days or even hours, and to take into account prematurity. For example, for the antibiotic gentamicin, which is commonly prescribed to neonates in the intensive care setting, one common guidelines is that a premature neonate weighing less than 1000 g at birth would get 3.5 mg/kg/dose every 24 h, but a term neonate less than 7 days old and weighing more than 1200 g would get 2.5 mg/kg/dose every 12 h, but the same infant over 7 days old (but less than 2000 g at birth) would get the same dose every 8–12 h (Taketomo et al. 2011). It's easy to see how such complex rules can be difficult to model in a prescribing system, and difficult to present to users in an intelligible way.

Off-Label Use Vendors of drug-dosing decision support databases, commonly used in EHR and e-prescribing products, offer guidelines for dosing that are used in decision support. Because many of the drugs used in pediatrics are not actually approved for use in children under a certain age, it can be seen as controversial for a vendor to provide a non-FDA-approved dose range. Because of the absence of FDA-approved dose ranges, local variation in these ranges is common. Such variation makes it even more difficult for drug-dosing decision support database vendors to provide this decision support confidently. The result is a database with incomplete data, where users of EHRs that use these data must fill in the blanks. Across data from multiple institutions, tremendous variation is seen in the dosing rules that are brought to bear on these prescriptions.

Metric vs. English Controversy Because of the dependency of changing body size on therapies, pediatric clinicians are in the habit of using metric-system measures for weight, height, temperature, and other measurements. Dosing guidelines are typically in milligrams (of drug) per kilogram (of patient's body weight) per day, and liquid concentrations are expressed in milligrams (of drug) per milliliter (of constituted drug product). The American public, however, has not taken up the metric system, so child health providers find themselves converting weights from kilograms to pounds, and doses of liquid medicines from milliliters to teaspoons. This conversion offers opportunity for error in the communication between physicians and families. It also offers a source of error in the data that is recorded for

prescriptions. Systems in an academic medical center will typically adhere carefully to metric units, but systems in community settings are more likely to store dosing guidelines and prescription records in terms of these imperial units. Merging data across sources must therefore take into account this conversion.

Compounded and Combined Medication Forms Owing to the inability of young children to swallow pills and the impracticality of the pharmaceutical market to provide liquid forms for all conceivable drugs, a small but significant number of medications must be converted to liquid form by a compounding pharmacy. Because the formulas for these compounded medications are not standard, the records for these drugs embedded in the EHR are not standard. Even within an institution, there can be multiple instances of compounding of the same medication that make comparison of prescribing data complex. Combination preparations, where more than one drug is in a preparation, are particularly common among compounded medications. Decision support aimed at one component of a combination medication may not be appropriate for the other components of the preparation, and users may be uncertain as to which component is the target of the decision support. The data model for the data in any analysis has to take these complexities into account.

Extra Requirements for Decision Support Rules Rules put in place to guide prescribing decisions in child health need to take body mass and age into account. As with any factor that makes decision rules more complex, the maintainability of the corpus of rules quickly outstrips the ability of any organization to maintain these rules. An effective general strategy for managing this complexity remains an unsolved conundrum (Conroy et al. 2007).

1.5 Immunization Management

1.5.1 Decision Support to Determine Currency of Immunizations

While adults receive immunizations according to schedules and risk factors, the complexity of the decision-making about which immunizations to give at what time is an order of magnitude greater. This is partly due to the higher number of targeted pathogens in immunization preparations, but also to the greater number of vaccine products on the market, the changing nature of vaccine guidelines, and the complexity of the temporal reasoning required for effective immunization administration. In addition, administration rules may change over time, presenting yet another challenge for accurate decision support. Below are the guidelines for administering the rotavirus vaccine. These rules illustrate the complexity that must be supported in an information system designed to give decision support for the administration of these

medications. It is also illustrates the common observation that some of the concepts used in the decision support are not computable from available data ("whenever possible," "clinically stable").

Rotavirus vaccine administration rules (Cortese et al. 2011)

- For Product 1, there are three required doses, to be given at 2, 4, and 6 months.
- For Product 2, there are two required doses, to be given at 2 and 4 months.
- The series should be completed with the same product whenever possible.
- The minimum age for administration of the first dose of either product is 6 weeks 0 days.
- The maximum age for administration of the first dose of either product is 14 weeks 6 days; if this deadline is missed, the vaccine should not be administered.
- The minimum interval between administrations of either product is 4 weeks.
- There is no maximum interval between doses, but the maximum age for administration of the final dose of either product is 8 months and 0 days
- Rotavirus vaccine may be give concomitantly with DTaP vaccine, HiB vaccine, IPV, hepatitis B vaccine, and pneumococcal conjugate vaccine
- Preterm infants can be immunized at the same schedule as term infants, based on the preterm infant's chronological (post-delivery) age. The preterm infant should be clinically stable, and should be given no earlier than the time of discharge from the neonatal intensive care unit or nursery.

1.5.2 Decision Support to Schedule Immunizations

The decision about what immunizations to deliver today is different from the decision about when the next doses are due. A convenient way to simplify this decision-making has evolved by way of the recommended schedule of well-child visits (Haggerty 1987). When new vaccines are introduced, their schedule conforms to this schedule in order to make it easier to administer. The rotavirus vaccine cited above, for example, conforms to the standard pattern of "well-baby checks" at 2, 4, and 6 months familiar to most parents. If a patient is able to stick to the prescribed schedule, there is little need for decision support to for what ought to be given at the visit and when to return for the next immunizations. Unfortunately, the real-life ability to adhere to this schedule is low (Selden 2006) so child-health providers are left with a great deal of decision making that they expect their information systems to help with. Only a minority of current EHRs do so (Kemper et al. 2006). This deficiency is due to the complexity of the logic required for sensible recommendations and the dependency on that logic on manually entered data.

1.5.3 Immunization Registries and Information Exchange

One solution to the problem of missing immunization data and recommendations for current or future doses lies in the technology of immunization registries, now more widely known as immunization information systems (IIS) (Yasuda and Medicine 2006). Because of the state-based organization of public health programs, and the state-based organization of the Medicaid program, these IIS are usually established to operate within a specific state or a region within a state. The state-based nature of these systems presents challenges to usability in populations that live near state borders. It also means that resources available to administrators of these systems are as constrained as any state-funded program (Blavin and Ormond 2011). The case for IIS is that since each patient typically receives immunizations in a variety of settings (doctors' offices, public health clinics, immunization fairs, schools, pharmacies) a unifying system will allow all providers to make decisions about who needs immunizations, and public resources can be directed to improve immunization rates in the highest risk areas. Standards exists for the transmission of immunization information (ref: HL7 v. 2.5.1 Implementation Guide for Immunization Messaging) but are usually customized in a way that makes interoperability between systems difficult. The U.S. federal Meaningful Use program (HHS 2009) at Stage 1 requires providers to conduct a test of data transmission to a public health agency; one choice in the menu of items aimed at this goal is the state immunization registry. While this is a far cry from requiring full interoperability, it is a step in the right direction toward encouraging the development of mature data sharing. In the meantime, providers who implement EHRs are faced with the dual challenge of getting local logic set up to support improvements in immunization rates while providing data to state IISs using manual methods and batch uploads.

1.6 Patient Identification

1.6.1 Newborn-Infant Transition

Newborn infants must be registered with a new medical record number (MRN) and a suitably distinctive name at the time of birth to allow normal clinical workflows. Typically these infants are assigned a name based on the mother's name, as is "Boy Jane Smith" or "Jane Smith Twin A." While the infant retains his or her MRN after this temporary name is changed to the child's given name, the MRN remains unchanged, but clinicians tend to assign more salience to a name than to a number. This change can make it challenging to integrate information across the perinatal period, especially when the venue of care changes (e.g. from the nursery to the doctor's office). EHRs used in the care of newborns must allow users to track patients based on either name. This is functionally similar to the tracking of adults who change their names after marriage, but in this case it happens to practically all

individuals. At stake in this identification process is the newborn screen results, sent in for analysis in the first few days of life, but whose results come back to the outpatient provider well after the baby's given name is established.

1.6.2 Fetal-Newborn Transition

The rise of techniques useful in the diagnosis and treatment of fetal problems presents special problems for healthcare record-keeping. Typically, patients do not receive a distinct record in the system until after they are born. Records maintained on fetal procedures are usually stored in the mother's chart, as is appropriate at the time of the procedure. Tying the information about fetal procedures to the record of the infant after birth requires at least a linkage between the mother's chart and the baby's. Ideally, there should be a way to split out information from the mother's chart into the infant's chart, in a way that preserves the usual privacy standards (just because it is clinically appropriate for one to access one person's chart does not mean it is appropriate to access another's). Currently in EHR systems this kind of fine access control is not technically possible. The clinician is left with the task of manually extracting information from the mother's chart into the baby's, pending development of systems that take fetal care into account.

1.6.3 Maternal-Fetal/Infant Linking

There is one circumstance where access to one person's chart entails some access to another's. In newborn care, there are elements of the mother's perinatal chart that are just as clinically relevant to the baby as the mother: Group-B Strep colonization status of the mother at birth, HIV status of the mother, medications given to the mother around the time of delivery, and so forth. While one can tie charts together in some EHR modules specifically designed for perinatal care, the movement of this information into the baby's chart in a way that would make this information extractable for analysis or available for decision support does not exist in general purpose EHR systems. Manual abstraction or a workaround using an external system is usually what is needed to support these data functions.

1.6.4 Pediatric-Adult Care Linking

Another instance where charts need to be linked across venues is when a pediatric patient "graduates" to adult care (Cooley 2011). Currently, from the information system perspective, the best practice is to transmitting a subset of the clinical data in the form of a care summary using the Continuity of Care Document (CCD)

format. In current technology there is no practical way to transmit the entire corpus of information on a particular patient.

1.7 Developmental Monitoring

A central feature of health supervision is screening for developmental delays. These delays in speech, motor function, or other abilities may indicate primary neurodevelopmental disease or be secondary to systemic disorders or socioeconomic factors. In any case, it is in the domain of the primary-care child health provider to screen for delays and to refer to needed services, like audiology, speech pathology, or neurology (ref: AAP Counc Child Disab 2006). The best practice for this activity is to use standardized developmental evaluation instruments—questionnaires—that can be filled out by the clinician or in some cases the parent. From the data perspective, the problems to be solved include how to marry parent-entered data into the medical record of the child, how to share developmental screening data for public health purposes, and how to track a very diffuse referral and follow-up process to ensure that no patients' needs go unaddressed. In addition, most of the developmental screening tools available are proprietary, which makes widespread implementation costly if not impossible.

1.7.1 Newborn Screening

Another screening process performed on practically all newborns in industrialized countries is newborn screening for congenital disorders. Often called "metabolic" screening, because of the emphasis on such metabolic diseases as phenylketonuria and hypothyroidism, most newborn screening programs now include screening for hearing loss (albeit not via a blood sample). Since these programs are state-run in the U.S., each state has a slightly different panel of disorders that it screens for. Challenges in the management of data for newborn screening include correct identification of the infant as he or she changes names, correct identification of the follow-up provider, presentation of the data in a way that can be stored and acted upon, and interoperability between state databases and provider EHRs. In the current environment, there is not widely implemented technical solution for any of these problems.

1.7.2 Well-Child Care

Applicable Guidelines The American Academy of Pediatrics has published recommendations for well-child care according to a schedule for many years (refs). State Medicaid programs expect that this schedule of visits (typically at birth, 1–2

weeks, 2, 4, 6, 9 and 12 months, then 15 and 18 months, then at 2 years and annually thereafter) be provided to their beneficiaries. The immunization schedule is arranged around the assumption of this schedule of visit, as are other guidelines for screening (anemia, developmental delay, lead poisoning, etc.). In addition to the timing of these visits, the AAP recommends what clinical events should occur at these visits: measurements, sensory screening, lab tests, and advice for parents. The Bright Futures guidelines (http://brightfutures.aap.org) provide more detail on the content of these visits.

Required Data Insurers and State Medicaid agencies set standards for the content of these well-child visits. In Medicaid programs, these requirements are embodied in the Early and Periodic Screening, Diagnosis, and Treatment program (EPSDT). Audits enforce these standards, but some agencies are requiring reporting of the actual data collected in these visits, such as the records of immunizations given, screening test results, and diagnoses. A messaging standard has been developed for this reporting (ref California EPSDT format). Quality measures for pediatric primary care are built on the assumption that this schedule of well-child visits is occurring.

1.8 Terminology Issues in Pediatric EHRs

Like all specialized areas of health care, pediatrics uses terminology differently from other areas. While there are special terms used in patient histories and exams, those terms general live in free-text portions of the record. Terminology specializations are most obvious in the regions of the EHR that focus on diagnoses (such as a problem list, past medical/family/social history, or billing section). For a diagnostic terminology system to be usable in child health, it must:

- Allow detailed descriptions of family historical risk factors
- Be descriptive of specific congenital syndromes and their subtypes
- Have detailed descriptors of anatomic anomalies that may lie outside of syndromes
- Allow description of chromosomal anomalies
- Describe patient behaviors that represent risk factors for undiagnosed behavioral disorders
- Describe family stressors that may affect child health and development
- Describe maternal conditions that affect the infant (e.g., “Maternal Group B Strep colonization”)
- Describe developmental, educational, or anthropometric states (none of which may be a disorder *per se*) throughout the lifespan

1.9 Pediatric Quality Measurement and the EHR

Quality measurement has been an important part of EHR implementation for adult care providers for many years (McColm and Karcz 2010) but the maturity of quality measures applicable to children is far behind that of adults. Part of this is due to the fact that outside of asthma and attention deficit hyperactivity disorder, chronic disease in children consists of small populations, in contrast to the large adult populations entailing diabetes, coronary artery disease, congestive heart failure, osteoporosis, and other high-prevalence adult conditions. For these numerous smaller populations, there are few simple proxy measures available to measure outcome, as one can do with diabetes (through the hemoglobin A1c percentage blood test) or process (as one can do with osteoporosis with data on the timing of bone-density studies). Nevertheless, there is increasing interest in pediatric measures that can be extracted from EHR data (Fairbrother and Simpson 2011; Kallem 2011) and the U.S. Federal Meaningful Use program is included more pediatric measures in its Stage 2 program (CMS 2012). As with any other measure, reliable data entry is prerequisite, so any data entry outside the normal workflow of ordering procedures, receiving lab test results, or prescribing medications is apt to be invalid. Special improvement programs aimed at data collection quality would usually be necessary to get to an acceptable level of validity if clinician data collection is expected.

Most quality measures are computed from claims data, owing to claims data's higher quality requirements as compared to EHR data. Those higher quality requirements are met in large part because the data requirements for claims are far simpler. The lack of detail in these simpler data sets fails to express the full complexity of care, so current efforts to develop more meaningful quality measures are taking EHR data into account, with some success (Angier et al. 2014; Bailey et al. 2014).

One phenomenon that tends to hamper the spread of pediatric quality measures is that all quality measures require sufficient numbers of patients to offer power to detect differences and to provide meaningful population estimates. While a large, tertiary children's hospital, may see enough patients with, say, testicular torsion in a year to afford a decent sample of patient outcomes data, the same is not true for general hospitals and community practices. It is therefore difficult to build momentum for quality measurement for most pediatric conditions. One recent study (Berry et al. 2015) noted that even among U.S. children's hospitals, few institutions had sufficient quantity of data to detect a drop in care quality for sickle cell disease, appendectomy, cerebrospinal fluid shunt surgery, gastroenteritis, heart surgery, and seizure of a reasonable amount over the 3-year timeframe of the study.

1.10 Registries and Population Management Within the EHR

The word *registry* can mean different things in the context of health information technology. Classically, a healthcare registry is a carefully curated set of data maintained for a specific analytical purpose, like tracking the current state of patients in

a geographic region with a particular chronic disease (e.g, cystic fibrosis (Sanders et al. 2015; Sykes et al. 2016)) or undergoing a particular kind of care (e.g., surgery for congenital heart disease (Husain et al. 2014; Pasquali et al. 2014)). In this form of registry, data from the electronic health record is usually extracted, reformatted to match the definitions of the registry, and uploaded through a standardized process. With the increasing prevalence of EHR systems in physicians' offices and hospitals, some curators of registries are crafting more direct, less expensive methods of identifying candidates for registry inclusion (Sarkar et al. 2014) or populating data directly into registries (e.g., The American Academy of Neurology's AXON registry (Goldenberg 2015)). Extracting data directly from the EHR places a burden on clinical users to assure high data quality (Kern et al. 2006); it remains to be seen whether direct connections to these kinds registries will truly obviate the labor-intensive data-formatting process of more traditional registry data loading. Inevitable variance between the data models of the EHR and the registry may require changes in either the EHR or the registry in order to facilitate seamless loading of data from one to the other (Merrill et al. 2013).

Registry in another sense refers to functionality found within the EHR system itself, in which patients are identified as being part of a group that is managed according to a plan of monitoring and outreach (Navaneethan et al. 2011). Typically, the inclusion and exclusion of a patient in a given registry is at least in part determined by criteria recorded in the EHR as part of routine care, like diagnoses. Once in the EHR registry, data on patients' disease state or risk stratification can be viewed to allow clinical workers to identify patients most in need of services. EHR registries typically facilitate the tracking of outreach activities (phone, email, letters, etc.) and the results of disease-modifying interventions. Decision support focused on the members of the registry can be implemented for only those patients, thereby helping to narrow alerts and reminders to the appropriate population. Because the registry provides a well defined denominator, the system can better compute meaningful measures of process and outcome, typically displayed in a dashboard-style display designed to drive clinical activities. Likewise, registry membership provides a validated way to identify populations for research studies and for the computation of metrics that require defined denominators, like immunization rates or measures of disease activity. Membership in a registry within the EHR is usually tracked with a data element (flag) that can be manipulated manually or computed from other data. Researchers using EHR data will need to use these flags to assure fidelity between clinical and research findings.

1.11 Conclusion/Summary

EHRs used in the care of infants, children, and adolescents must support different functionalities than systems designed for adult care. Adaptations to these systems to accommodate pediatric clinical work will affect the type of data available. Those who seek to use data from pediatric EHRs should examine how well the specialty

workflow is supported by the EHR in order to be able to interpret the system's output. Use of pediatric healthcare data for secondary uses such as quality reporting and registries must take all of these factors into account in order to be effective.

References

AHRQ. Agency for healthcare research and quality children's electronic health record format. 2013. Retrieved from https://healthit.ahrq.gov/health-it-tools-and-resources/childrens-electronic-health-record-ehr-format.

Andrews AL, Kazley AS, Basco Jr WT, Teufel 2nd RJ. Lower rates of EMR use in rural hospitals represent a previously unexplored child health disparity. Hosp Pediatr. 2014;4(4):211–6.

Angier H, Gold R, Gallia C, Casciato A, Tillotson CJ, Marino M, Mangione-Smith R, DeVoe JE. Variation in outcomes of quality measurement by data source. Pediatrics. 2014;133(6):e1676–82.

Bailey LC, Mistry KB, Tinoco A, Earls M, Rallins MC, Hanley K, Christensen K, Jones M, Woods D. Addressing electronic clinical information in the construction of quality measures. Acad Pediatr. 2014;14(5 Suppl):S82–9.

Berry JG, Zaslavsky AM, Toomey SL, Chien AT, Jang J, Bryant MC, Klein DJ, Kaplan WJ, Schuster MA. Recognizing differences in hospital quality performance for pediatric inpatient care. Pediatrics. 2015;136(2):251–62.

Blavin F, Ormond B. HITECH, meaningful use, and public health: funding opportunities for state immunization registries. Washington, DC;2011. Retrieved from http://www.medicaidhitechta.org/Portals/0/Users/011/11/11/ImmunRegWhitePaper.pdf.

Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med. 2010;363(6):501–4.

CCHIT. CCHIT child health workgroup. 2010. Retrieved from http://www.cchit.org/workgroups/child-health.

CDC. Percentile data files with LMS values. 2012. Retrieved from http://www.cdc.gov/growthcharts/percentile_data_files.htm.

CMS (Center for Medicare and Medicaid Services). Medicare and Medicaid: Electronic Health Records (EHR) Incentive Programs. 2012. Retrieved January 16, 2016, from https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/.

Conroy S, Sweis D, Planner C, Yeung V, Collier J, Haines L, et al. Interventions to reduce dosing errors in children: a systematic review of the literature. Drug Saf. 2007;30(12):1111–25.

Cooley WC, Sagerman PJ. Supporting the health care transition from adolescence to adulthood in the medical home (Transitions Clinical Report Authoring Group). Pediatrics. 2011;128(1):182–200.

Cortese MM, Leblanc J, White KE, Jerris RC, Stinchfield P, Preston KL, Meek J, Odofin L, Khizer S, Miller CA, Buttery V, Mijatovic-Rustempasic S, Lewis J, Parashar UD, Immergluck LC. Leveraging state immunization information systems to measure the effectiveness of rotavirus vaccine. Pediatrics. 2011;128(6):e1474–81.

Dart RC, Rumack BH. Intravenous acetaminophen in the United States: iatrogenic dosing errors. Pediatrics. 2012;129(2):349–53.

Fairbrother G, Simpson LA. Measuring and reporting quality of health care for children: CHIPRA and beyond. Acad Pediatr. 2011;11(3 Suppl):S77–84.

Fenton TR. A new growth chart for preterm babies: Babson and Benda's chart updated with recent data and a new format. BMC Pediatr. 2003;3:13.

Fenton TR, Kim JH. A systematic review and meta-analysis to revise the Fenton growth chart for preterm infants. BMC Pediatr. 2013;13:59.

Finnegan B, Ku L, Shin P, Rosenbaum S. Policy research brief no. 9: boosting health information technology in medicaid: the potential effect of the American Recovery and Reinvestment Act.

2009. Retrieved from http://www.gwumc.edu/sphhs/departments/healthpolicy/dhp_publications/pub_uploads/dhpPublication_506602E1-5056-9D20-3D7DD946F604FDEE.pdf.
Friedlin J, Dexter PR, Overhage JM. Details of a successful clinical decision support system. AMIA Annu Symp Proc. 2007; 254–8.
Gardner RM, Pryor TA, Warner HR. The HELP hospital information system: update 1998. Int J Med Inform. 1999;54(3):169–82.
Goldenberg JN. The breadth and burden of data collection in clinical practice. Neurol Clin Pract. 2015;10–1212.
Green M. 50 things to know about the EHR market's top vendors. 2015. Retrieved from http://www.beckershospitalreview.com/healthcare-information-technology/50-things-to-know-about-the-ehr-market-s-top-vendors.html.
Grummer-Strawn LM, Reinold C, Krebs NF, (CDC), C. f. D. C. a. P. Use of World Health Organization and CDC growth charts for children aged 0-59 months in the United States. MMWR Recomm Rep. 2010;59(RR-9):1–15.
Haggerty RJ. Health supervision visits: should the immunization schedule drive the system? Pediatrics. 1987;79(4):581–2.
HHS (U.S. Department of Health and Human Services). Health Information Technology for Economic and Clinical Health (HITECH) Act, Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA). Pub. L. No. 111–5, 123 Stat. 115, 516 (Feb. 19, 2009).
Husain SA, Pasquali SK, Jacobs JP, Hill KD, Kim S, Kane LC, Calhoon JH, Jacobs ML. Congenital heart operations performed in the first year of life: does geographic variation exist? Ann Thorac Surg. 2014;98(3):912–8.
Kallem C. Transforming clinical quality measures for EHR use. NQF refines emeasures for use in EHRs and meaningful use program. J AHIMA. 2011;82(11):52–3.
Kemper AR, Uren RL, Clark SJ. Adoption of electronic health records in primary care pediatric practices. Pediatrics. 2006;118(1):e20–4.
Kern EF, Maney M, Miller DR, Tseng CL, Tiwari A, Rajan M, Aron D, Pogach L. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. Health Serv Res. 2006;41(2):564–80.
Kim G, Lehmann C. Pediatric aspects of inpatient health information technology systems. Pediatrics. 2008;122(6):e1287–96.
Kirkendall ES, Spooner SA, Logan JR. Evaluating the accuracy of electronic pediatric drug dosing rules. J Am Med Inform Assoc. 2014;21(e1):e43–9.
Kressly S. Testimony of Susan Kressly, MD, FAAP, practicing pediatrician, American Academy of Pediatrics, before the Committee on Small Business, Subcommittee on Regulations and HEalthcare, U.S. House of Representatives. 2009.
Laron Z, Lilos P, Klinger B. Growth curves for Laron syndrome. Arch Dis Child. 1993;68(6):768–70.
Lehmann CU, O'Connor KG, Shorte VA, Johnson TD. Use of electronic health record systems by office-based pediatricians. Pediatrics. 2015;135(1):e7–15.
Leu MG, O'Connor KG, Marshall R, Price DT, Klein JD. Pediatricians' use of health information technology: a national survey. Pediatrics. 2012;130(6):e1441–6.
Martin ND, Smith WR, Cole TJ, Preece MA. New height, weight and head circumference charts for British children with Williams syndrome. Arch Dis Child. 2007;92(7):598–601.
McColm D, Karcz A. Comparing manual and automated coding of physicians quality reporting initiative measures in an ambulatory EHR. J Med Pract Manage. 2010;26(1):6–12.
Merrill J, Phillips A, Keeling J, Kaushal R, Senathirajah Y. Effects of automated immunization registry reporting via an electronic health record deployed in community practice settings. Appl Clin Inform. 2013;4(2):267–75.
Miller RA, Waitman LR, Chen S, Rosenbloom ST. The anatomy of decision support during inpatient care provider order entry (CPOE): empirical observations from a decade of CPOE experience at Vanderbilt. J Biomed Inform. 2005;38(6):469–85.

Nakamura MM, Ferris TG, DesRoches CM, Jha AK. Electronic health record adoption by children's hospitals in the United States. Arch Pediatr Adolesc Med. 2010;164(12):1145–51.

Navaneethan SD, Jolly SE, Schold JD, Arrigain S, Saupe W, Sharp J, Lyons J, Simon JF, Schreiber Jr MJ, Jain A, Nally Jr JV. Development and validation of an electronic health record-based chronic kidney disease registry. Clin J Am Soc Nephrol. 2011;6(1):40–9.

Nellhaus G. Head circumference from birth to eighteen years. Practical composite international and interracial graphs. Pediatrics. 1968;41(1):106–14.

ONC. Comprehensive list of certified health information technology. 2016. Retrieved from http://oncchpl.force.com/ehrcert.

Pasquali SK, Jacobs ML, He X, Shah SS, Peterson ED, Hall M, Gaynor JW, Hill KD, Mayer JE, Jacobs JP, Li JS. Variation in congenital heart surgery costs across hospitals. Pediatrics. 2014;133(3):e553–60.

Ranke MB, Heidemann P, Knupfer C, Enders H, Schmaltz AA, Bierich JR. Noonan syndrome: growth and clinical manifestations in 144 cases. Eur J Pediatr. 1988;148(3):220–7.

Rollins JD, Collins JS, Holden KR. United States head circumference growth reference charts: birth to 21 years. J Pediatr. 2010;156(6):907–3, 913.e901–2.

Rosenbloom ST, Qi X, Riddle WR, Russell WE, DonLevy SC, Giuse D, Sedman AB, Spooner SA. Implementing pediatric growth charts into an electronic health record system. J Am Med Inform Assoc. 2006;13(3):302–8.

Sanders DB, Fink A, Mayer-Hamblett N, Schechter MS, Sawicki GS, Rosenfeld M, Flume PA, Morgan WJ. Early life growth trajectories in cystic fibrosis are associated with pulmonary function at age 6 years. J Pediatr. 2015;167(5):1081–8. e1081.

Sarkar IN, Chen ES, Rosenau PT, Storer MB, Anderson B, Horbar JD. Using Arden Syntax to identify registry-eligible very low birth weight neonates from the Electronic Health Record. AMIA Annu Symp Proc. 2014;2014:1028–36.

Selden TM. Compliance with well-child visit recommendations: evidence from the Medical Expenditure Panel Survey, 2000–2002. Pediatrics. 2006;118(6):e1766–78.

Shiffman RN, Spooner SA, Kwiatkowski K, Brennan PF. Information technology for children's health and health care: report on the Information Technology in Children's Health Care Expert Meeting, September 21-22, 2000. J Am Med Inform Assoc. 2000;8(6):546–51.

Spooner SA. Special requirements of electronic health record systems in pediatrics. Pediatrics. 2007;119(3):631–7.

Spooner SA, Classen DC. Data standards and improvement of quality and safety in child health care. Pediatrics. 2009;123 Suppl 2:S74–9.

Strom BL, Schinnar R, Aberra F, Bilker W, Hennessy S, Leonard CE, Pifer E. Unintended effects of a computerized physician order entry nearly hard-stop alert to prevent a drug interaction: a randomized controlled trial. Arch Intern Med. 2010;170(17):1578–83.

Sykes J, Stanojevic S, Goss CH, Quon BS, Marshall BC, Petren K, Ostrenga J, Fink A, Elbert A, Stephenson AL. A standardized approach to estimating survival statistics for population-based cystic fibrosis registry cohorts. J Clin Epidemiol. 2016;70:206–13.

Taketomo CK, Hodding JH, Kraus DM. Lexi-Comp's pediatric & neonatal dosage handbook: a comprehensive resource for all clinicians treating pediatric and neonatal patients (Pediatric dosage handbook). 18th ed. Hudson: Lexi-Comp; 2011.

Yasuda K, Medicine, C. o. P. a. A. Immunization information systems. Pediatrics. 2006;118(3):1293–5.

Chapter 2
Protecting Privacy in the Child Health EHR

S. Andrew Spooner

Abstract In the United States and other industrialized countries, laws demand that all individually identifiable health information be secured from unintended disclosure and handled as private, sensitive information. While this protection extends equally to all information in a health record, information that pertains to mental health, reproductive health, physical abuse, and certain other areas with social impact is usually considered even more sensitive than other types of health information. The latter types of information may have special laws or professional standards that apply to how it is handled. All of these privacy and security issues become more complex in situations where minors are involved, because of real or perceived conflicts between the interests of the child and the interests of parents or guardians. In the care of adolescents, these issues become particularly difficult, and may affect how data are recorded or displayed in the EHR system, and the extent to which data may be available for research. Additional areas that present difficult challenges to privacy include fetal care, foster care, and situations where genetic information must be stored and interpreted. Security policies for access to systems intended to be used by patients (personal health records and patient portals) are complex. They can become even more challenging when the child has participated in clinical research and unexpected clinically relevant results are obtained. In this chapter we will discuss the prevailing regulations in the United States and the European Union that apply to privacy and security, and highlight pediatric aspects of these rules that apply to data.

Keywords Data integrity • Privacy • Security • HIPAA

S.A. Spooner, M.D., M.S., FAAP (✉)
Departments of Pediatrics and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine,
3333 Burnet Avenue, MLC-9009, Cincinnati, OH 45229, USA
e-mail: andrew.spooner@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_2

2.1 The Information in an EHR

2.1.1 Basic EHR Data Integrity

Like the research record, the data in the electronic health record demands a high level of integrity. While the goal of research record data integrity policies is to ensure scientific rigor, the goals of EHR data integrity are of a different nature:

- Unlike a research record, the EHR tends to be inclusive of all data--not just validated data. For example, a health record may contain two blood pressure recordings at an office visit because the clinician decided to repeat the measurement. One would not expect the original recording to be deleted. In a case where "the" encounter blood pressure needs to be recorded for research, the researcher has the dilemma of what to do with such an inclusive data set.
- Whereas data collection for research typically follows rigid and well-defined data collection processes, this is not so for the clinical health record. The result is that a clinical record is much less structured than a case report form. Automated extraction of data for research is therefore challenging.
- The data in the EHR belong to the patient, and must be provided to the patient or the applicable guardian at any time. The patient can also request changes to the chart (although these requests do not have to be honored if they are inappropriate) and the patient/parent may also add documentation to the chart at any time under HIPAA (OCR 2002) in the United States.
- The EHR plays an important role in legal defense of malpractice claims. Although the medical record is classified as hearsay (Elias 2003), one may still use it in court if one can show that the record is maintained in a businesslike way. Any evidence that the medical record is being used for purposes other than clinical care may render the record useless in legal defense. For this reason, there are usually limitations on which people in which job roles are allowed to make entries in the record.

2.1.2 Data Entry

Clinical care is documented typically by the recording of a large amount of free text and a small amount of discrete data. While EHRs can vary in the extent to which they demand discrete data entry, it is accepted that free-text entry (in the form of dictation, text-generating macros, or typing) is necessary to capture the complexity of clinical care. One might be able to reduce very simple patient encounters to a series of check-boxes with associated discrete data elements, but in academic medical centers where even moderately complex disease is addressed, it is not reasonable to expect clinicians to adhere to templates that generate primarily discrete data.

There are areas of the EHR, like laboratory test results and medication orders, that do contain a preponderance of discrete data, but there are some limitations to the uses of these data for research. In these areas there are usually a number of regulatory agencies that govern how these data are structured. For example, U.S. clinical laboratory procedures are certified through a program (CMS 2014; Kroger 1994) by federal law. Under these conditions, one is not free to set up investigational clinical laboratory tests as a part of routine care. Likewise, prescription data must conform to data standards that allow electronic prescribing (Liu et al. 2011), so investigational drugs present a challenge to represent in clinical EHRs. For example, if the investigational medication is not yet on the market, and therefore is not yet assigned a code from the system used to identify retail products, it might be difficult or impossible to encode as discrete data in the EHR. These regulatory hurdles, while they serve a good purpose, may make it impossible to use the EHR itself as a research record, even if the proper institutional review board assurances are obtained. "Shadow records" that parallel the clinical record for research can cause confusion in the clinical operation, especially when the research activities overlap with normal clinical activities.

Another particular challenge of maintaining research data that parallels clinical data is how to handle discrepancies between the two. It is customary to apply data quality standards to research data. For example, one may want to select a particular blood pressure, collected under certain conditions, for a data point in a research study. One might then delete all other blood pressures from the research record in order to establish the data point of interest. This kind of deletion of data is not possible in an EHR, assuming the data were not collected in error. All data are retained, and deleting data—even if it is erroneous—must be done in a way that retains the data for future inspection. Most clinical operations that allow corrections of data in the EHR have strict policies about how the change is documented. It would be unusual to see a situation where data from a clinical research study would flow back to the clinical record as a correction, regardless of how valid the correction might be. In any case, only those personnel authorized to make entries in the clinical record can initiate those changes.

2.2 Privacy Concepts in Pediatrics

Health care information is sensitive, and as such is protected by the Administrative Simplification provisions of the Health Insurance Portability and Accountability Act of 1996 (Chung et al. 2006), as well as state laws. Because every episode of pediatric care involves at least two people in a patient role (the patient and the patient's parent or guardian), and perhaps many more, the task of securing information while maintaining the appropriate level of access is especially challenging in pediatrics. As technology moves toward fulfilling the goal of faster information flow and higher transparency, these issues are exacerbated. Pediatric clinical research,

especially in genomics, can also generate health care information that creates privacy concerns. These issues are discussed in Chap. 6, Data Governance and Strategies for Data Integration and Chap. 7, Laboratory Medicine and Biorepositories.

2.2.1 HIPAA

The U.S. Health Insurance Portability and Accountability Act of 1996 intended to provide for continuity of health insurance coverage after a change in employer. This initial goal never materialized, but the portion of the law that required electronic transmission of health care claims (Title 2) remained. This portion of the regulation, known as "Administrative Simplification," raised concerns about privacy and security of the claims information that was required to be sent. This concern spawned the HIPAA Privacy Rule and the HIPAA Security Rule, enacted in April 2003 and currently enforced by the U.S. Office of Civil Rights (HHS 2002). While the full detail of these rules is beyond the scope of this text, it is important to appreciate that HIPAA remains the main driver of how clinicians behave to protect health information (privacy rules, mostly) and how systems are designed to protect it (security). An important principle regarding the use of EHR information is the "minimum necessary" rule, which states that those who access the record see only that part of the record that is necessary for performance of their job. This principle affects (or should affect) users' behavior, but it also guides policies for who is given access to what parts of the EHR. A researcher wanting to examine records of patients solely for the purposes of research would violate this rule. The HITECH Act of the American Recovery and Reinvestment Act of 2009 (HHS 2013) strengthen HIPAA's privacy and security requirement, and impose stiffer penalties.

2.2.2 HIPAA Business Associate Agreements

Those who work with health data, unless the data are suitably rendered anonymous, are subject to the HIPAA privacy and security rules, and the attendant penalties, through business associate agreements. These agreements bind recipients of health care data to the same rules that the clinical originators of data must follow, and applies the same penalties for breaches of confidentiality. Recent changes in US law regarding business associates (2013) have reinforced the seriousness of the government in its intent to enforce these rules.

2.2.3 Pediatric Aspects of HIPAA

The HIPAA Privacy Rule allows parents or guardians access to the child's health information in almost all situations. Exceptions include when the minor is the one who consents to care and the consent of the parent is not required under State or

other applicable law; or when the minor obtains care at the direction of a court; or if the parent agrees that the minor and the health care provider may have a confidential relationship (HHS 2002). Privacy laws vary from state to state, and providers are obliged to follow the most stringent one. Since control of children's health information is sometimes a hot political topic (as in the case of minors' access to reproductive health services) these legal conflicts can make control of data very complicated (Chilton et al. 1999).

2.2.4 FERPA

A law that existed many years before HIPAA was the Family Educational Rights and Privacy Act (Kiel and Knoblauch 2010) which attempts to give students some control over the use of their educational records. When healthcare is provided at a school, the line between health records and educational records is blurred, and there can appear to be conflicts between HIPAA and FERPA. If one is attempting to aggregate data from both educational and healthcare settings, these specific laws may come into play. The U.S. Department of Education and the U.S. Department of Health and Human Services published joint guidance on navigating these apparently conflicting laws in 2008 (HHS 2008).

2.2.5 Release of Information

A common function of the information systems in a healthcare organization is the release of information based on a request from a patient, parent, guardian, lawyer, State agency, or other suitably approved group. Release of information (ROI) in a hospital is typically handled via a controlled process through the Health Information Management department or similar entity. Before the age of EHRs, the actual conveyance of medical records was achieved by a tedious and time-consuming process of photocopying paper records or printing images of documents from archived storage. It was considered normal for this process to take several weeks. The difficulty of this process rendered the medical record effectively inaccessible to all, but the most dedicated patients and their representatives.

In the information age, expectations about the ease by which one can get information are changing. The Continuity of Care Document project (Ferranti et al. 2006) is a manifestation of the expectation that electronic health records can produce immediate summary information for the purposes of sharing across venues of care. The expectation of immediate access has spread to all areas of the EHR (How et al. 2008). These expectations entail more sophisticated authentication methods than the typical notarized permission form that usually initiates the process of ROI today.

ROI is important to understand in pediatric care because it means that all information in the chart (or at least that part designated the "legal medical record") is available to the guardian at all times. While it may have been comforting to assume

that the information is "secure" from prying parental eyes because of a 6-week wait for photocopying, that wait will eventually be reduced to practically zero through electronic methods. Parents or guardians will have contemporaneous access to all details in a child or adolescent's chart. We have not yet had the opportunity to evolve habits in practice that take this into account, or sophisticated privacy policies that balance the need to keep things truly private between a provider and a minor patient under the assumption of immediate parental electronic access.

2.2.6 Clinical Data Sharing vs. Financial Data Sharing

Regardless of privacy policies put in place, the fact that guardians receive billing information about health services provided also runs counter to the concept of keeping things private between a minor and a provider. Doctors who treat adolescents have been known to write prescriptions on paper or provide samples rather than run the risk of notifying a parent via a pharmacy claim. Regardless of how one feels about the appropriateness of such confidential care, such practices do create holes in the protections set up in the electronic record.

2.2.7 Parental Notification vs. Consent to Treat

Adolescents can consent to treatment at an age younger than the age of majority in certain clinical contexts (Weddle and Kokotailo 2002). For example, an adolescent at age 12 can, in the states of California or Illinois (as of 2003 (English and Kenney 2003; Kerwin et al. 2015)) consent to treatment for mental health services. In North Carolina, the minor can consent at any age. This varying age of consent has little impact on EHR functionality or data storage, but it is often confused with the concept of parental notification. Just because an adolescent can consent to treat for his or her own care does not make the record of that treatment confidential, or obviate parental notification regulations. Once again, the availability of that information in the medical record may appear threatening to both patient and provider, to the point that the provider may record data in a non-standard place (like a "sticky note" field that is not part of the legal medical record). Once again, full appreciation of the workflow used to produce health data is necessary in order to construct meaningful queries and analysis.

2.2.8 Mandated Reporting

Child health workers are obliged under the law of all U.S. states to report suspected child abuse. This obligation overrides HIPAA or other concepts of health information privacy (AAP 2010).

2.2.9 The European Data Protection Directive

In the European Union, the right to privacy and its effect on data management is reflected in Directive 95/46/EC, commonly known as the Data Protection Directive (DPD) of 1995 (Barber and Allaert 1997), and the subsequent 2012 proposed reforms of this law (Saracci et al. 2012). The scope of this directive is larger than health care, but does apply to EHR data. The focus of these laws is to ensure protection of inter-country transfer of information as part of clinical care, but any inter-country use of data, including research, would be affected. Of course, within-country handling of data would be governed by laws within that nation. In many European countries (e.g., Denmark, Finland), there are centralized database of health information, but the use of these databases for research is controversial (Lehtonen 2002) and the DPD does not specifically address how data might be used in medical research (Sheikh 2005). Similarly, adolescent privacy is not specifically addressed in the DPD, although it is reasonable to assume that the laws of individual countries would take precedence. As the "right to be forgotten" legislation from the European Union (Jones 2016) indicates, European privacy laws that might apply to medical data may be even more restrictive than in the United States. It is unclear whether this focus in privacy will work for or against an adolescent's interest, since parents' interests and the adolescents' interest can be in conflict.

2.3 Health Information Privacy in Adolescent Care

2.3.1 The Nature of Adolescent Practice

The care of adolescent patients—as in the care of all patients—must address issues of particular sensitivity: reproductive health, sexually transmitted disease, substance abuse, physical abuse, eating disorders, sexual abuse, mental health, and sexual orientation (Gray et al. 2014). The difference with adolescents that affects EHR implementation is that the patients are more sensitive to the effects of confidentiality on their decision to seek care (Ginsburg et al. 1997; Ginsburg et al. 1995). Most agree that adolescents need to share in the decision-making about their care, regardless of their inability to legally consent to their treatment. For sensitive topics, adolescents may forego care in order to hide information from parents (Britto et al. 2010; Ford et al. 2004). Since a fundamental goal of health information technology is usually to make information *easier* to share, the adolescent's prerequisite to restrict information dissemination may be impossible to accommodate without the establishment of special policies and procedures. As a result, clinical users may resort to obfuscation of data or the use of paper to manage the information that would otherwise be contained in the EHR. Obviously, this would have major downstream effects on the interpretation of data derived from these environments.

2.3.2 *Data Access Policies in the Adolescent Patient*

Adolescent Health, Privacy and Research Adolescents participate as subjects in clinical research, but the process for weighing the risks and benefits of parental consent are complex. Even when parental consent is not a sensitive issue, researchers intending to engage in clinical research involving adolescents should familiarize themselves with local legal issues regarding assent and consent at various ages. The Society for Adolescent Health and Medicine maintains guiding policies for these issues (Bayer et al. 2015; Santelli et al. 1995).

Confidential Care It is a basic principle of adolescent healthcare, endorsed by professional societies, that they be offered confidential care when appropriate (ACOG 2014; Ford et al. 2004; Gans Epner 1996). Since health information is already considered confidential, a promise of confidential care essentially means that information will be kept from parents or guardians, a concept that flies in the face of some state law and EHRs designed to provide information to parents or guardians in the form of printed summaries and on-line portals. As of this writing, there are no standards for adolescent privacy policies to govern such patient-accessible information, whether for clinical care or research.

2.4 Health Information Privacy and Mental Health

Mental health information was singled out in the HIPAA Administrative Simplification rules in the sense that "psychotherapy notes" do not have to be disclosed to patients or families as part of the usual release of information. These kinds of notes are usually made to record a therapist's thoughts during a patient's therapy, and, if a patient accessed these notes, they might be damaging to the patient's progress. The regulation specifies that these notes cannot contain "medication prescription and monitoring, counseling session start and stop times, the modalities and frequencies of treatment furnished, results of clinical tests, and any summary of the following items: diagnosis, functional status, the treatment plan, symptoms, prognosis, and progress to date" (HHS 2001).

This minor exception to the idea that a patient or family owns the information in the chart with complete access rights has no direct effects on data analysis. It does, however, impose requirements for more complex access control on developers of EHRs. It also has the potential to confuse clinical users, who are already struggling with how to practice medicine in the era of patients' immediate access to their information. For example, if psychotherapy notes should not be shared, are there not other classes of data in the chart that ought to be afforded this same protection, for the same reasons? HIPAA did not describe other exceptions, but clinicians' desire to document care without disrupting care may create new use cases that make data access policies even more complex than they are now.

2.5 Guardianship Issues (Adoption, Foster Care, Fetal Care)

In pediatrics, as with elder care, the patient is not assumed to be the main decision-maker in health care decisions. For most children, the parents are responsible for the child's care as well as the financial and administrative transactions involved in that care. In some cases, the guardian must be distinguished from the financial guarantor. For children whose parents have had their parental rights severed, or who have otherwise been taken from the care of their parents, other adults are designated guardians. In specific legal proceedings, a court may appoint a *guardian ad litem* with defined decision-making authority for the child. The only impact these complex arrangements may have on data used for research is that it may affect the consent processes associated with the study.

References

AAP. American academy of pediatrics, committee on child abuse and neglect policy statement: child abuse, confidentiality, and the health insurance portability and accountability act. Pediatrics. 2010;125(1):197–201.

ACOG. American college of obstetrics and gynecology, committee opinion no. 599: committee on adoscent health care: adolescent confidentiality and electronic health records. Obstet Gynecol. 2014;123(5):1148–50.

Barber B, Allaert FA. Some systems implications of EU data protection directive. Stud Health Technol Inform. 1997;43 Pt B: 829–833.

Bayer R, Santelli J, Klitzman R. New challenges for electronic health records: confidentiality and access to sensitive health information about parents and adolescents. JAMA. 2015;313(1):29–30.

Britto MT, Tivorsak TL, Slap GB. Adolescents' needs for health care privacy. Pediatrics. 2010;126(6):e1469–76.

Chilton L, Berger JE, Melinkovich P, Nelson R, Rappo PD, Stoddard J, Swanson J, Vanchiere C, Lustig J, Gotlieb EM, Deutsch L, Gerstle R, Lieberthal A, Shiffman R, SA Spooner Stern M. American academy of pediatrics. Pediatric practice action group and task force on medical informatics. Privacy protection and health information: patient rights and pediatrician responsibilities. Pediatrics. 1999;104(4 Pt 1):973–7.

Chung K, Chung D, Joo Y. Overview of administrative simplification provisions of HIPAA. J Med Syst. 2006;30(1):51–5.

CMS. Center for Medicare and Medicaid Services, CLIA program and HIPAA privacy rule; patients' access to test reports. Final rule. Fed Regist. 2014;79(25):7289–316.

Elias CE. Federal rules of evidence handbook. Durham: Carolina Academic Press; 2003.

English A, Kenney K. State minor consent laws: a summary. 2nd ed. Chapel Hill: Center for Adolescent Health and the Law; 2003.

Ferranti JM, Musser RC, Kawamoto K, Hammond WE. The clinical document architecture and the continuity of care record: a critical analysis. J Am Med Inform Assoc. 2006;13(3):245–52.

Ford C, English A, Sigman G. Confidential health care for adolescents: position paper for the society for adolescent medicine. J Adolesc Health. 2004;35(2):160–7.

Gans Epner JE. Policy compendium on reproductive health issues affecting adolescents. Chicago: American Medical Association; 1996.

Ginsburg KR, Slap GB, Cnaan A, Forke CM, Balsley CM, Rouselle DM. Adolescents' perceptions of factors affecting their decisions to seek health care. JAMA. 1995;273(24):1913–8.

Ginsburg KR, Menapace AS, Slap GB. Factors affecting the decision to seek health care: the voice of adolescents. Pediatrics. 1997;100(6):922–30.
Gray SH, Pasternak RH, Gooding HC, Woodward K, Hawkins K, Sawyer S, Anoshiravani A. Society for adolescent health and medicine: recommendations for electronic health record use for delivery of adolescent health care. J Adolesc Health. 2014;54(4):487–90.
HHS. Does the HIPAA Privacy Rule allow parents the right to see their children's medical records? 2002;03/14/2006. Retrieved from http://www.hhs.gov/hipaa/for-professionals/faq/227/can-i-access-medical-record-if-i-have-power-of-attorney/.
HHS. Joint guidance on the application of the Family Educational Rights And Privacy Act (FERPA) and the Health Insurance Portability And Accountability Act of 1996 (HIPAA) to student health records. Washington, DC. 2008. Retrieved from http://www2.ed.gov/policy/gen/guid/fpco/doc/ferpa-hippa-guidance.pdf. Accessed 4/02/2012.
HHS (U.S. Department of Health and Human Services). Title 45 – Public Welfare (October 1, 2001). 45 C.F.R. pt. 164.
HHS (U.S. Department of Health and Human Services). Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification rules under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; other modifications to the HIPAA rules. Fed Regist. 2013;78(17), 5565–702.
How SKH, Shih A, Lau J, Schien C. Public views on U.S. health system organization: a call for new directions, vol. 11. New York: The Commonwealth Fund; 2008.
Jones ML. Ctrl+Z: the right to be forgotten. London: New York University Press; 2016.
Kerwin ME, Kirby KC, Speziali D, Duggan M, Mellitz C, Versek B, McNamara A. What can parents do? A review of state laws regarding decision making for adolescent drug abuse and mental health treatment. J Child Adolesc Subst Abuse. 2015;24(3):166–76.
Kiel JM, Knoblauch LM. HIPAA and FERPA: competing or collaborating? J Allied Health. 2010;39(4):e161–5.
Kroger JS. Coping with CLIA. Clinical laboratory improvement amendments. JAMA. 1994;271(20):1621–2.
Lehtonen LA. Government registries containing sensitive health data and the implementation of EU directive on the protection of personal data in Finland. Med Law. 2002;21(3):419–25.
Liu H, Burkhart Q, Bell DS. Evaluation of the NCPDP structured and codified sig format for e-prescriptions. J Am Med Inform Assoc. 2011;18(5):645–51.
OCR (Office of Civil Rights, U.S. Department of Health and Human Services). Standards for privacy of individually identifiable health information.67 Fed.Reg. 53181 (August 14, 2002) (to be codified at 45 CFR pts. 160 & 164).
Santelli JS, Rosenfeld WD, DuRant RH, Dubler N, Morreale M, English A, Rogers AS. Guidelines for adolescent health research: a position paper of the society for adolescent medicine. J Adolesc Health. 1995;17(5):270–6.
Saracci R, Olsen J, Seniori-Costantini A, West R. Epidemiology and the planned new Data Protection Directive of the European Union: a symposium report. Public Health. 2012;126(3):253–5.
Sheikh AA. The Data Protection (Amendment) Act, 2003: the Data Protection Directive and its implications for medical research in Ireland. Eur J Health Law. 2005;12(4):357–72.
Weddle M, Kokotailo P. Adolescent substance abuse. Confidentiality and consent. Pediatr Clin North Am. 2002;49(2):301–15.

Chapter 3
Standards for Interoperability

S. Andrew Spooner and Judith W. Dexheimer

Abstract Semantic interoperability between clinical information systems is a major goal of current pediatric IT implementations and ongoing research. For both clinical care and research collaboration, successful exchange of meaningful clinical data depends on flexible, standard formats for the construction of messages, and widely accepted terminologies to capture the clinical concepts. In 2016, most messages take the form of delimited character strings that adhere to the Health Level 7 (HL7) version 2.X standard. We review the state of the art of version 2 HL7 messaging types and describe a well known example of this kind of messaging in the CDC Implementation Guide for Immunization Messaging. Version 3 of HL7, built on XML, has the potential for richer semantics but is not yet widely used. We use the Continuity of Care Document as a pediatric example of the use of this standard in real systems. Terminology systems, both open and proprietary, are used to encode clinical and administrative concepts in pediatrics. We will review terminology systems in current use and their pediatric-specific limitations, and mention some current efforts to create a platform for applications that interact with EHRs 18 (SMART) with the messaging standard that supports it (FHIR).

Keywords Standards • Direct secure messaging • Terminology • HL7 • SNOMED • ICD-10 • SMART • FHIR

S.A. Spooner, M.D., M.S., FAAP (✉)
Departments of Pediatrics and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine,
3333 Burnet Avenue, MLC-9009, Cincinnati, OH 45229, USA
e-mail: andrew.spooner@cchmc.org

J.W. Dexheimer, Ph.D.
Departments of Pediatrics and Biomedical Informatics, Divisions of Emergency Medicine and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Ave, ML-2008, Cincinnati, OH 45229, USA
e-mail: judith.dexheimer@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_3

3.1 Introduction

Semantic interoperability—that is, the sharing of data between EHR systems where clinical meaning is preserved—is a major goal of governments, the EHR industry, and the field of medical informatics (ONC 2016). Because clinical data is always the product of workflow, and because workflows differ between environments, between specialties, and even between providers, maintaining clinical meaning is challenging. Add the fact that software used for the same purpose is usually implemented very differently across vendors and even across customers within the same software developer's installed base, and it is no wonder that semantic interoperability of health data has been described as the holy grail of medical informatics (Benson 2012). Data interchange and analysis would certainly be easier if all information systems used the same data model. Despite the development of the Health Level 7 Reference Information Model (Rocca et al. 2006; Shakir 1997) realities of software implementation preclude such uniformity, so the informatics community must concentrate its standards-development activities on the format of data transmissions between systems (messaging) and the semantic payload contained in the transmission (terminology). In this chapter, we will review the basics of messaging standards used in health information technology, and review terminology systems of importance to pediatric care.

3.2 Standards Development Organizations and Messaging Standards

Voluntary standards development organizations (SDOs) aim to gather consensus on the structure of messaging standards that can be employed for practical use cases.

3.2.1 ANSI

The American National Standards Institute accredits standards and the organizations that produce them. It coordinates the efforts of voluntary SDOs and sets standards for behavior to assure openness in processes and robustness of product. There is no pediatric-specific subgroup within ANSI.

3.2.2 HL7

Health Level 7 (after level 7 of the Open Systems Interconnection model) is the most widely recognized health-related standards development organization in the world. Its messaging standards are used commonly in the data interchange between

clinical systems, like laboratory systems and electronic health record systems. There is a Child Health Workgroup within HL7 (2012). HL7 produces a wide variety of standards, including some that are not messaging standards. For example the Arden Syntax for Medical Logic Systems (Anand et al. 2015; Hripcsak et al. 2015; Samwald et al. 2012), is a standard way to represent medical decision-making rules, including the required inputs, the expected outputs, the text of alerts, and embedded orders. The aforementioned Reference Information Model (Rocca et al. 2006; Shakir 1997) is not itself a messaging standard, but it guides the data structure of all artifacts generated within the HL7 universe of standards.

HL7 messages are complete units of data intended to be transmitted between systems in a single transmission. Segments within the message contain logical groupings of data fields. A field contains up to a defined number of characters that comply with the definition of a data type (string, integer, etc.). Depending on the version of HL7 you are using, fields can be delimited as logical units within an XML message (version 3) or merely delimited by the pipe ("|") character in a character string (version 2). Figure 3.1 illustrates the contrast between a version 2 and version 3 message.

An implementation guide is a user manual intended to instruct system programmers on how to construct compliant messages. HL7 International publishes these guides in conjunction with subject matter experts and organizations. For example, the HL7 Version 2.5.1 Implementation Guide for Immunization Messaging (Savage and Williams 2014) describes message formats for requesting immunization records, for receiving information on candidate patient patches from a query, immunization administration records, and administrative reporting related to vaccine adverse events. This guide is a collaboration between HL7, the U.S. Centers for Disease Control, and the American Immunization Registry Association.

HL7 Version 2 This version is widely implemented in all healthcare settings. It is a much simpler (but more loosely defined) model for messaging than version 3, consisting of definitions of strings that can be assembled on the sending end and parsed on the receiving end into semantically appropriate components. The HL7 version 2 messaging standard allows for locally-defined "Z" segments not defined by any HL7 standard. Use of Z segments leads to localizations of messages that make them impossible to use without special customizations on the receiving end.

HL7 Version 3 and the Clinical Document Architecture A more modern messaging standard developed by HL7 in 2000 is the Clinical Document Architecture or CDA (Dolin et al. 2001, 2006; Goossen and Langford 2014). It is more "modern" in that it is based on a reference information model—specifically the HL7 Reference Information Model—and development of document definitions follows a standard modeling process. CDA document definitions are expressed in XML, so specimens that conform to these definitions can use XML document-type definitions to present multiple views, including human-readable ones. In a sense, a CDA document is always more than a message, since it is, by definition, a complete clinical document that exists in a clinical context. One advantage of CDA is that it

HL7 Version 2 Message Segment

```
OBX|1|SN|1554-5^GLUCOSE^POST 12H CFST:MCNC:PT:SER/PLAS:QN||^182|mg/dl|70_105|H|||F<cr>
```

HL7 Version 3 XML Fragment

```
<observationEvent>
 <id root="2.16.840.1.113883.19.1122.4" extension="1045813"
     assigningAuthorityName="GHH LAB Filler Orders"/>
 <code code="1554-5" codeSystemName="LN"
       codeSystem="2.16.840.1.113883.6.1"
       displayName="GLUCOSE^POST 12H CFST:MCNC:PT:SER/PLAS:QN"/>
 <statusCode code="completed"/>
 <effectiveTime value="200202150730"/>
 <priorityCode code="R"/>
 <confidentialityCode code="N"
    codeSystem="2.16.840.1.113883.5.25"/>
 <value xsi:type="PQ" value="182" unit="mg/dL"/>
 <interpretationCode code="H"/>
 <referenceRange>
    <interpretationRange>
        <value xsi:type="IVL_PQ">
        <low value="70" unit="mg/dL"/>
        <high value="105" unit="mg/dL"/>
      </value>
      <interpretationCode code="N"/>
    </interpretationRange>
 </referenceRange>
```

Fig. 3.1 Contrast between versions 2 and 3 of HL7 messaging standards. In each case below, the portion of the message shown is intended to convey a blood glucose value (182 mg/dL) obtained 12 h after a stress test. The message includes the normal expected range of 70–105 mg/dL. The version 3 fragment is much longer, but embeds much richer semantics. The object identifiers (e.g., "2.16.840.1.113883.19.1122.4") are standard index numbers of other standards embedded in the message, like specific terminology systems

supports incremental interoperability, insofar as implementers can start with a simple document, with appropriate semantic tags, and still exchange data while waiting to implement more complex semantics as the application matures. Later in this chapter we will examine an example of CDA, the Continuity of Care Document.

3.2.3 Committee E31 on Healthcare Informatics of ASTM International

ASTM (2016) (formerly American Society for Testing and Materials) promulgates standards relevant to health care. The guides produced by ASTM tend to be less clinical and more administrative, as in standards for passing transcribed text or securing transmitted health information. There is no pediatric-specific subgroup within ASTM International. ASTM and HL7 collaborated on the Continuity of Care Document, described below.

3.2.4 Accredited Standards Committee (ASC) X12

This group develops messaging standards for electronic data interchange of administrative data in a variety of industries, including healthcare (ASC 2016). ASC X12 standards are not very clinical and have no particular pediatric flavor, but researchers may see these standard messages used in conjunction with clinical systems for transmitting insurance and payment information.

3.2.5 The National Council for Prescription Drug Programs (NCPDP)

Computer systems have proven patient safety utility in the area of medication prescribing (Johnson et al. 2013; Kaushal et al. 2010). For electronic prescribing to reach its full potential, there needs to be a standard way of exchanging prescriptions between providers and pharmacies (among other actors). NCPDP, a not-for-profit, ANSI-accredited SDO produces a suite of standards that allow transmission of data related to pharmacy operations, including the SCRIPT standard for prescriptions, a federally endorsed standard (CMS 2010; Liu et al. 2011). While the EHR may store its medication orders and prescription data in its own data model, the required standard for transmitting the prescription to the pharmacy is the SCRIPT standard, which forces the EHR to format the data to satisfy the requirements for constructing a valid SCRIPT message.

3.3 Trends Toward Interoperability

3.3.1 Meaningful Use

The Health Information Technology for Economic and Clinical Health (HITECH) act within the 2009 American Recovery and Reinvestment Act (HHS 2009) called for the appropriation of federal funds to stimulate the adoption of electronic health records among U.S. health care providers. To ensure that the systems that providers implement were actually fully-functional EHRs, there was a requirement that these systems be certified that they could perform certain functions. There was another requirement that providers be able to demonstrate that they were using the systems "meaningfully"—in other words, in a way that might conceivably benefit the patient. Several provisions of this "meaningful use" (MU) rule entailed the use of standards, like electronic prescribing and exchanging data with other providers. An unfortunate side effect of MU was that compliance with MU, which arguably voluntary, absorbed much of the resources of providers and vendors that might otherwise have been spent on innovation and responsive design. Penalties for not complying with

MU (part of the Medicare program that adult providers, but not pediatric providers, would participate in) raised legitimate questions as to whether the program could be said to be truly voluntary. In any case, the initiative definitely increased the adoption of health information technology and the concomitant use of data standards by child health providers (Lehmann et al. 2015; Nakamura et al. 2015). As of this writing, the MU program is still in effect, although interest in it is waning across all specialties as incentive money dries up.

3.3.2 Direct Messaging

Fundamental to the goal of semantic interoperability is the ability to send secure messages containing clinical information directly from one clinical provider to another. Today, in the United States, most such communication occurs by fax, since there is no widely implemented, uniform method of addressing recipients or of ensuring security of electronic forms of communication. The Direct Project (ONC 2014; Sujansky and Wilson 2015) establishes the protocols by which trusted agents known as Health Information Service Providers (HISPs) manage the secure messaging backbone for these messages. Occasionally the project is referred to as the DIRECT project (Sujansky and Wilson 2015), but the word is not an acronym. HISPs all over the United States are working out how providers should communicate these messages to them, using a uniform message-addressing system also specified by these standards. Of course, HIPAA standards for privacy and security apply to these communications, but to the extent they do, Direct offers the first reasonable way for data to move between electronic health record systems. At first, most of the messages will simply be messages intended to be read by human recipients, but plans are in place for allowing structured data to be carried in these messages. Reports exist in the literature citing the rapid rise in the use of Direct messaging to populate EHRs and patient portals (Reicher and Reicher 2016).

3.3.3 Electronic Prescribing

In Stage 1 of the U.S. MU program, providers were required to send at least 30 % of prescriptions via electronic prescribing; for stage 2, it was 50 %. While it does not affect pediatricians, the Medicare Improvements for Patients and Providers Act of 2008 (MIPPA) (CMMS 2008) requires most adult providers to adopt e-prescribing as well. Because most providers are adult providers, this has ensured the rapid spread e-prescribing as the most common method of providing prescriptions in the United States. The opportunities for rich datasets for analysis under these conditions will be revolutionary.

3.3.4 Quality Reporting Programs

The rise of electronic health records comes with an increased interest in direct, electronic reporting of quality measures. While it does not affect child health providers (since they see no Medicare patients outside the narrow context of end-stage renal disease) the Physician Quality Reporting Initiative (PQRI) incentive program as part of Medicare has already stimulated the creation of techniques for data extraction from health information systems. Between MU and the Child Health Insurance Program Reauthorization Act (CHIPRA) of 2009, child health providers are also getting into the business of direct quality measure reporting. Stage 2 of the MU program included a significant number of pediatric-relevant quality measures for electronic reporting. The net effect of these programs is that providers will have to adopt systems that have more robust data management capabilities than before. As a secondary result, they will be more likely to want to use those capabilities for analysis outside of mandated reporting.

3.3.5 Personal Health Records and Consumer Empowerment

Prior to the widespread implementation of EHRs, information, stored on paper, was difficult to move. It was laborious to package up copies of papers for other providers, so, in some cases, providers wrote letters to each other, summarizing care. Patients or families who wanted access to their health information were similarly hobbled by the process of photocopying, but were often willing to create an independent summary record of health care, in such systems as the now defunct Google Health or the currently active Microsoft Health Vault (Do et al. 2011). Such untethered personal health record systems are themselves difficult to maintain without unusual levels of dedication. As providers use computers more and more, patients will value summary information from these systems. Health care providers will provide information from their office systems to share with personal health records, or at least provide a patient portal into a subset of the provider's system. The need to combine records across very diverse health care providers will encourage the adoption of standards for both moving the data and retaining the meaning of these records as they move into different contexts.

3.4 HL7 Examples

3.4.1 HL7 v2 Example: Immunization Data Messaging (CDC Implementation Guide)

Within child health, the task of maintaining statewide immunization information systems (formerly known as registries) is the most important inter-institutional data exchange task. The implementation guide for immunization messaging (Savage and

```
RXA|0|1|20120401|20120401|03^MMR^CVX|.5|ML^^ISO+|||987654321^DOE^JOHN^G^^^^^^^^^VEI~123123123^O=DOKES^JOSEPH^A^^DR^
MD^^^^^^OEI|^^^PRIMARY CARE CLINIC^^^^^100 SPRING STREET^^CINCINNATI^OH|||W3213213321|20120603|MSD^MERCK^MVX
RXR|SC^SUBCUTANEOUS^HL70162|LA^LEFT ARM^HL70163
```

Fig. 3.2 In this fragment of an HL7 version 2 message, an immunization registry is sending a record of an immunization. In real life, this message would be "wrapped" in data indicating the identity of the patient, the sender, and the receiver. RXA is the segment that contains a record of the administration of the vaccine. Within the RXA segment there are items such as RXA-5, which identified the substance administered (measles-mumps-rubella vaccine, in this case, encoded from a standard terminology system), and RXA-11, the location of the organization where the administration occurred. Other data include the identity of who administered the vaccine, the lot number, and the expiration date. The RXR segment describes the anatomical site to which the vaccine was delivered, in this case the *left arm*

Williams 2014) written by the American Immunization Registry Association (AIRA) and the Centers for Disease Control and Prevention (CDC) is therefore the most clinically relevant example of real-world pediatric data exchange. This implementation guide lays out how messages containing information on patients, their immunization history, and immunization events (administrations, reactions, forecasts), without constraining how the EHR handles immunization data storage or decision support internally. Figure 3.2 illustrates the structure of a version 2 message.

3.4.2 HL7 v3 Example: Continuity of Care Document

Fundamental to consumer empowerment in health care is consumers' access to summary information like diagnoses, medications, allergies, immunizations, and past medical events. While the record contains much more that just these summary data, much of the remaining data has limited usefulness beyond its original temporal context. These data are also useful to other providers, of course. Base on a paper form used in inter-hospital transfer in Massachusetts, the Continuity of Care Record project was born, and grew into an electronic message format defined by ASTM (Ferranti et al. 2006). Eventually this message standard was harmonized with the HL7 CDA and became the Continuity of Care Document (CCD). CCD is the basis for exchanging summary records in at least some electronic systems. The expectation is that all electronic systems will be able to read and write CCDs. Desire for accurate CCDs (accurate because providers will get feedback from patients and other providers if there are errors) will drive processes for accurate data on which to base them, which can only be good for the state of patient data in electronic systems. Figure 3.3 illustrates the structure of a portion of a CCD.

```
problemObservation{
    id(root:'9d3d416d-45ab-4da1-912f-4583e0632000')
    code(code:'ASSERTION', codeSystem:'2.16.840.1.113883.5.4')
    value{
        builder.build{
            cd(code:'494627013',codeSystem:'2.16.840.1.113883.6.96',displayName:'Hyaline Membrane Disease')
        }
    }
    problemStatus{
        value(code:'413322009', codeSystem:'2.16.840.1.113883.6.96',displayName:'Resolved')
    }
    problemHealthstatus{
        value(code:'162467007', codeSystem:'2.16.840.1.113883.6.96',displayName:'Symptom Free')
    }
```

Fig. 3.3 Fragment of XML from a continuity of care document. The above XML expresses a resolved problem of hyaline membrane disease, which you might find in a Continuity of Care Document from a newborn intensive care admission. SNOMED-CT is indicated by the use of the object identifier (OID) 2.16.840.1.113883.6.96

3.5 Terminology Standards

3.5.1 Definitions

Terminology Systems Terminology systems are collections of related concepts defined within a discipline—are intended to convey unambiguous meaning within a defined context. Many terminology systems exist as defined standards, maintained by standards-development organizations, which specify the applicable domain, the meaning of the words, the relationships between the words, the mapping to codes, and use cases within the clinical domain. Unfortunately, in real clinical systems, system implementers react to user dissatisfaction with the rigidity of using a tightly defined list of terms by opening the list up to local customizations. Those who must grapple with data extracts from EHRs need to clarify to what extent a terminology standard is being used in its pure form, or even within its intended domain. The most important reason to adhere to the pure form of a terminology is to support interoperability—both between people and between machines.

Terminology systems exist for a variety of clinical domains, like diagnoses, procedures, exam findings, organisms, anatomic terms, psychosocial problems, and other clinical concepts. To the extent that system implementers use the appropriate terminology for the right part of the application, terminologies can facilitate *semantic interoperability*; that is, the meaning of the record can be conveyed across organizational boundaries. This kind of data exchange is a major goal of the U.S. effort to create a National Health Information Network (Dixon et al. 2010).

Post-coordination vs. Atomic Concepts When designing the terms for a terminology system that a human user will see, developers have the dilemma of how much coordination of concepts to do. If one does no pre-coordination (bringing together of separate concepts into a more complex concept, as in “fracture of the humerus,” where “fracture” and “humerus” are the atomic [indivisible] concepts) then users will be left with the laborious task of bringing them together themselves. If one does too much pre-coordination, then users may not be able to find the exact concept they want to express (e.g., “exacerbation of mild intermittent asthma” may

be an inappropriate term for a child with exercise-induced asthma with exacerbation). For terminology systems that are attached to common, real-world processes, the level of coordination may be straightforward (e.g., in LOINC, the term 57698-2 refers to "lipid panel with direct LDL," a heavily pre-coordinated term, but one that represents a laboratory test that is actually obtained in practice). Clearly, pre-coordination imposes a tremendous maintenance burden on the terminology developer, since there are new, complex concepts created every day in clinical medicine. It also creates an issue with historical data, since a given pre-coordinated term might not have been available at the time the data were collected. A classification system that groups related concepts is often necessary when aggregating terms that were collected via clinical processes over many years. All terminology systems, no matter how pre-coordinated, undergo periodic updates, of course, but the more clinically relevant the terminology is, the more rapidly it will be updated. These updates present special challenges for data aggregation, as mentioned, but in pediatric applications there are apt to be more local terminology customizations to accommodate special pediatric terms not found in standard terminologies.

Types of Terminology Systems Sets of terms are usually designed with a specific purpose in mind. An *interface terminology system* is made to support data entry (the "interface" here is the human-computer interface). As such, an interface terminology system emphasizes the use of natural language terms with synonyms. Terms in these systems are heavily pre-coordinated, in order that they may be selected rapidly for real clinical situations. SNOMED-CT is an example of an interface terminology system. In contrast, a *reference terminology system* is designed for data processing like retrieval, aggregation, and determination of equivalent data points. Reference terminology systems may appear to a clinical user to be overly rigid, but this is the result of a reference terminology system's main functional requirement to eliminate ambiguity. A *classification* is a set of terms intended to divide concepts into groups for a defined purpose. The most familiar classification system is the International Classification of Diseases (e.g., ICD-9, Clinical Modification, used to encode diagnoses for all U.S. healthcare claims), maintained by the World Health Organization. A common blunder in system implementation is to use a classification like ICD-9-CM when an interface terminology system would much better capture the concepts that clinical users are trying to express. Interface terminology systems typically map their terms to an appropriate classification so that it can be captured for administrative purposes, allowing the user to use the more appropriate reference information system.

3.5.2 Pediatric Aspects of Terminology Systems

Terminology used to describe pediatric concepts must match the findings, conditions, and treatments found in the pediatric environment, of course. But there are some requirements of pediatric terminology systems that go beyond the simple

description of clinical phenomena. In addition to very specific descriptions of diseases seen only in pediatrics, terminology systems must also accommodate the capture of data where terms may need to apply to the patient and the parent separately. For example, a term like "38-week gestation" may refer to the mother's pregnancy or the child's time of delivery. Likewise, terminologies must capture the fluid nature of children's growth and development, and social situations that apply only to the young. For example, descriptors of a patient's capability of self care may represent an abnormal state in adults but a normal state in infants.

Diagnoses and Conditions The most familiar use of a standard terminology system in pediatric medicine is for diagnoses, including descriptions of conditions for which no diagnosis is (yet) known. Physicians are usually very familiar with the assignment of diagnosis codes in the billing process. Since the terminology system used for billing (the classification ICD-9- ICD-9-CM) is not especially rich, physicians will often incorrectly assume that computer systems cannot adequately capture clinical concepts with adequate specificity. In fact, there are several terminology systems that clinicians can use that do an excellent job of capturing clinically relevant concepts. The dilemma the system implementer has is which one to use, and whether to try to ask users to use a reference terminology system directly, or a more clinically relevant interface terminology.

3.6 Terminology Systems

3.6.1 SNOMED-CT (The Systemized Nomenclature of Medicine-Clinical Terms)

SNOMED-CT is a extensive clinical terminology standard formed in 2002 when the College of American Pathologists merged the England and Wales National Health Service's Clinical Terms (a UK-based terminology for primary care previously known as the Read Codes) (Harding and Stuart-Buttle 1998) with the reference terminology subset of SNOMED known as SNOMED-RT (Cornet and de Keizer 2008). The resulting system, called SNOMED-CT (for Clinical Terms) is intended to be used as a general-purpose reference terminology system, including the domains of diagnoses, conditions, historical findings, exam findings, and test results. SNOMED began as a pathology terminology system, it has expanded its domain to the point that the National Library of Medicine (NLM) licensed it for use throughout the United States in 2003 (Cornet and de Keizer 2008). Since 2007, SNOMED-CT has been maintained by the International Health Terminology Standards Development Organisation (IHTSDO) with the goal of promoting international adoption of its use. One can obtain direct access to SNOMED-CT by obtaining a license to use the NLM's Metathesaurus (https://www.nlm.nih.gov/databases/umls.html).

3.6.2 Unified Medical Language System

The goal of any terminology system, including SNOMED-CT, is to reduce variability in data capture and encoding in order to facilitate better clinical care and research (Ruch et al. 2008). Unfortunately, no terminology system has perfect coverage of all conceivable concepts, so the natural evolution of terminology systems in clinical care is to customize the content, or develop home-grown terminologies. Mapping such homegrown systems to standard terminologies usually shows that the reference terminology system fails to cover all the needed concepts by a significant margin (Rosenbloom et al. 2008; Wade et al. 2008). The U.S. National Library of Medicine's Unified Medical Language System's (UMLS) Metathesaurus is an attempt to draw lines of semantic equivalence between terms of various systems, but is not itself used as a clinical terminology system. The Metathesaurus is, as the name implies, more of a thesaurus than a terminology system. Via the UMLS Terminology Services web site (https://uts.nlm.nih.gov/) one can access SNOMED-CT and other terminologies that are mapped within UMLS.

SNOMED-CT is a good example of a terminology system that includes a semantic network, indicating relationships between concepts. The simplest romantic relationship is the "is a" relationship, as in "von Gierke disease **is a** glycogen storage disease" (a network of terms connected by "is a" relationships is also known as a *taxonomy*). In this relationship, "glycogen storage disease" is the parent of the child concept, "von Gierke disease." Semantic networks can be useful in EHRs, but their navigation adds complexity to the user interface that limits their use. For example, a user may want to take a concept that is in the record and express it with more specificity. Following "is a" links to child concepts can achieve this specificity, provided that the EHR embeds those links. Semantic relationships can be more sophisticated, and can, for instance, connect diseases to the symptoms that it causes or a drug to its common side effects. Other relationships include "is caused by," "site of finding," and "manifests as." An EHR equipped with semantic links between terms can give significant decision support to a user who, for example, may want to generate a list of symptoms commonly associated with a disease "on the fly" for documentation purposes. The structure of SNOMED CT contains the current concept, parents and children of that concept and descriptors of the concept including defining relationships, qualifiers, and descriptions or synonyms. It is freely available for research use and can be accessed both online and through the Internet (http://snomed.dataline.co.uk/).

SNOMED-CT can be used within pediatrics; adaptation and extension of the terms has been performed in a research setting to help improve pediatric care and concepts (James et al. 2009).

3.6.3 ICD

International Classification of Disease (ICD) classifications (ICD-9-CM, ICD-10) encompass many of the same domains as SNOMED-CT (diagnoses, symptoms, and procedures), but are not intended to be an interface terminology system. ICD-10's larger number of terms may make it seem to users that ICD is becoming more usable as an interface terminology system for diagnoses. While it is true that the higher granularity of ICD-10 will allow better use of administrative (claims) data in clinical decision support, ICD is still a classification system, so system implementers should continue to use a true interface terminology system to capture these data.

ICD-9 vs. ICD-190 ICD-10 contains approximately 70,000 codes for diagnoses compared to ICD-9's 14,000. Many (~48,000) of these new codes are simply due to the introduction of the concept of which phase of care the patient is in (initial,follow-up, or sequelae) and 25,000 of the new codes simply add laterality (left, right, bilateral) to the associated term. Eleven thousand terms merely add the concept of phase of bone healing for fractures. Even so, ICD-10 does add some more detail to many existing diagnoses, making these codes arguably more useful for both clinical care and data analysis. The Clinical Modification of ICD-9, in use in the U.S. since 1979 (Slee 1978) is entrenched in datasets, so it will be years before the new code set, mandated in the U.S. since October 1, 2015, can be used without reference to its predecessor. Unfortunately, the mapping from ICD-9 to ICD-10 is not one-to-one or even one-to-many, so there is no way to automatically map older codes to new ones. An apt pediatric example is the reconfiguration of the codes for asthma. The old classification in ICD-9 of asthma into extrinsic, intrinsic, chronic obstructive, and other is now replaced by mild intermittent, mild persistent, moderate persistent, and severe persistent types. While this reclassification is a welcome modernization of the terminology on asthma, it makes managing a dataset spanning the transition difficult (Slee 1978). Particularly unfortunate is the observation that the extra detail afforded by ICD-10 does very little to add detail necessary for recording important features of pediatric disease. For example, ICD-10 allows recording of whether an ear infection is on the left of the right—a detail of dubious importance in almost any research project. And yet ICD-10 does not include detail on which mitochondrial myopathy a patient has—a detail that would be of much more scientific value. EHRs have implemented systems to help clinicians navigate this unfamiliar territory through decision support

3.6.4 LOINC

Logical Observation Identifiers Names and Codes (LOINC) is a freely available "universal code system for laboratory and clinical observations" developed in 1994 by the Regenstrief Institute and Indiana University (Huff et al. 1998). The database includes medical and laboratory codes, nursing diagnoses, nursing interventions,

outcome classifications, and patient care data. The laboratory coding in LOINC includes chemistry, hematology, serology, microbiology, and toxicology. The clinical portion of the coding includes vital signs, intake/output, procedures, selected survey instruments and other clinical observations. LOINC's primary use is to identify laboratory tests. It is endorsed by the American Clinical Laboratory Association as well as the College of American Pathologists. Since laboratory systems used in child health settings use certified clinical labs, LOINC supports the coding of tests used in pediatrics.

3.6.5 SMART and FHIR

Initiatives on implementing new healthcare data exchange standards, such as the Fast Healthcare Interoperability Resources (FHIR) framework (HL7 2015; Kasthurirathne et al. 2015a, b), have been introduced to standardize data across EHRs. The FHIR framework as an emerging standard for the interoperable exchange of EHR data, contains previous standards from HL7 and an application programming interface (API) for exchanging data. Going hand in hand with FHIR is a project known as Substitutable Medical Applications and Reusable Technologies (SMART), a standards-based, open-source technology platform designed to create applications that share data across EHR resources including warehouses, health information exchanges, EHRs, and any other repository of data in the healthcare system. "SMART on FHIR" (Alterovitz et al. 2015; Mandel et al. 2016) aims to integrate applications with EHRs and other health information data warehouses, and is gaining popularity in the research world applications. The organizers of the SMART project offer a gallery of working applications that have been deployed with real HER systems (SMART 2016).

The impact of an app-based health IT system on the data for clinical research is as yet unknown. The main argument for the SMART platform is that it will allow more flexibility in functionality and user interface, which arguably could support more productive data-capture processes, or promote more robust health information exchange (Bender and Sartipi 2013). The FHIR standard, in draft form as of this writing, has been embraced by all major health IT vendors as a way to help solve the interoperability challenge.

3.6.6 Other Terminology Systems

There are many other terminology systems in use in pediatric applications, like terms from the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders, Fifth edition (APA 2013) or the American Medical Association's Current Procedural Terminology (Abraham et al. 2011). Outside the United States, ICD-10 has been used for several years to classify diagnoses, and

SNOMED-CT has gained traction as a reference terminology since it incorporated the Read codes and is maintained by an international agency. Of course, interface terminologies, which must reflect the language and clinical practice of real end users, will continue to be especially heterogeneous across countries.

3.6.7 Proprietary Systems

3.6.7.1 MEDCIN

MEDCIN® is a proprietary system of standardized vocabulary developed by Medicomp Systems, Inc. (Chantilly, VA, http://www.medicomp.com/products/medcinengine/) for use as an interface terminology system for electronic medical records. Introduced in 1978, it now includes more than 280,000 clinical concepts, linked to each other by a rich semantic network. MEDCIN terms describe chief complaints, symptoms, exam findings, diagnostic test findings, and diagnoses, among others. The semantic network can be implemented within an EHR to allow one to generate a template of symptoms associated with a possible disease on the fly, for example. The terms in MEDCIN are mapped to terms in reference terminology systems and classifications like SNOMED-CT, LOINC, and ICD.

3.6.7.2 Intelligent Medical Objects

Intelligent Medical Objects (IMO, Northbrook, IL, http://www.e-imo.com/) develops and maintains medical terminology systems for problems (symptoms, diagnoses), procedures, and medications. IMO's emphasis is on providing pre-coordinated terms that a clinician in practice would normally use, allowing clinicians to use clinically relevant descriptors rather than the exact code necessary to achieve some administrative goal. The result of this approach is there are many synonymous terms, which vary from each other only in minor lexicographic features. As with any heavily pre-coordinated terminology system, there is much ongoing development, and the terms are updated continuously. Customers can request new codes based on new clinical phenomena. Using a dataset coded in IMO will necessarily involve referring to the crosswalk to standard terminology systems, or consulting with a clinician about nuances of meaning between similar terms. IMO is particularly useful in pediatrics due the terminology's ability to cover rare disease found primarily in pediatric practice.

3.6.7.3 Others

There are other terminology systems available, like the commercial products offered by Health Language (Denver, CO, http://www.healthlanguage.com/) and the public domain Omaha System, used chiefly by nurses in classifying patient problems,

interventions, and outcomes. In all cases, expertise about the intended use of the codes is required in order to use these systems in data analysis. Appropriateness for pediatric applications varies among these systems and in certain clinical domains within each one.

3.7 Medications

3.7.1 RxNorm (http://www.nlm.nih.gov/research/umls/rxnorm/index.html)

Standardized vocabularies also exist for medication-specific uses. While one could use straightforward chemical descriptions to name medications (like 2-acetoxybenzoic acid for aspirin) these standard names are not used in real clinical practice, and are absurdly complex to boot (e.g., (2S,5R,6R)- 6-([(2R)-2-amino-2-(4-hydroxyphenyl)- acetyl]amino)- 3,3-dimethyl- 7-oxo- 4-thia-1-azabicyclo[3.2.0]heptane- 2-carboxylic acid for amoxicillin). The National Library of Medicine embarked upon development of a standard naming terminology, RxNorm, that attempts to assign a unique concept identifier for each synonymous drug term complied from several drug databases. The databases from which RxNorm draws terms are sold commercially or published by the U.S. government. The avowed purpose of RxNorm is to provide drug names that describe medications that patients actually receive, which explains the approach of aggregating real-world drug databases. The technique RxNorm uses is to identify non-conflicting codes from within the combined data, and to assign a concept identifier. RxNorm's data model includes semantic relationships between concepts like "has-ingredient" for combination preparations or "has-tradename" to identify brands. Because RxNorm is derived from databases that primarily comprise products that are sold, it is lacking in content for drugs that are prepared ad hoc by compounding pharmacies. This deficit has profound implications for pediatrics, where there are children who must be treated with medications that are not available commercially, like a compounded liquid preparation derived from a crushed pill ordinarily used in adults. Another challenge to pediatric coding of medications is in compounded combination preparations for topical drugs. These recipes are not standardized. RxNorm is free and there is a publicly available browser (http://rxnav.nlm.nih.gov/).

3.7.2 Proprietary Medication Databases

Medi-Span Medi-Span is maintained by Wolters Kluwer Health (Medi-Span, Indianapolis, IN). Like other commercial drug databases, Medi-Span includes not just a coding system for medications but also clinical content regarding drug interactions, allergens existing in drug products, acceptable dose ranges, and other clini-

cal information needed for appropriate prescribing. It is this information that represents the main value that these databases offer. None of these databases are specific to pediatrics.

First Databank First Databank (FDB) is a clinical drug and drug interaction database designed and maintained by First Databank, Inc (South San Francisco, CA).

Multum Multum is a drug database maintained by Cerner Corporation (Kansas City, MO) and is embedded with the electronic medical record systems and healthcare services they provide, but can also be implemented in non-Cerner systems.

Other Commercial Databases There are other, similar commercial databases (e.g., Gold Standard' Elsevier, Netherlands) which implementers might choose. The wide diversity of these coding systems presents challenges to data aggregation, but RxNorm promises to help. In some EHR systems, implementers can choose which commercial drug database to embed in the system. In others, the EHR application is tied to a particular database. EHRs that are otherwise identical, then, can be fundamentally non-interoperable based on the variance in the coding scheme used to represent drug concepts. It is this variance that RxNorm is intending to rectify, but the technical overhead of mapping to standard terms is not yet commonly undertaken.

3.8 Conclusions

There are other standards operating in healthcare, like the imaging standards set by the Digital Imaging and Communications in Medicine (DICOM) standards for diagnostic imaging. To the extent that data standards for terminology and messaging achieve national scope, they will necessarily be focused on the needs of large, national populations—namely, adults. It is unlikely that any pediatric-specific data standard would gain enough traction to achieve widespread use. For this reason, pediatric informaticians recognize that advocating for the needs of children within larger data standards is a more productive than trying to make systems that specialize in pediatric needs. Researchers who work with data that employ these standards need to be aware that standards implemented in the pediatric healthcare environment are at high risk for non-standard implementations because of the differences between pediatric and adult care.

References

Abraham M, Ahlman JT, Anderson C, Boudreau AJ, Connelly JL. CPT 2012 (Cpt/Current Procedural Terminology (Professional Edition)). Chicago: American Medical Association Press; 2011.

Alterovitz G, Warner J, Zhang P, Chen Y, Ullman-Cullere M, Kreda D, Kohane IS. SMART on FHIR genomics: facilitating standardized clinico-genomic apps. J Am Med Inform Assoc. 2015;22(6):1173–8.

Anand V, Carroll AE, Biondich PG, Dugan TM, Downs SM. Pediatric decision support using adapted Arden Syntax. Artif Intell Med. 2015.

APA. American Psychiatric Association: diagnostic and statistical manual of mental disorders. 5th ed. Arlington: American Psychiatric Association; 2013.

ASC. ASC X12. 2016. Retrieved from http://www.x12.org/.

ASTM. Committee E31 on healthcare informatics. 2016. Retrieved from http://www.astm.org/COMMITTEE/E31.htm.

Bender D, Sartipi K. HL7 FHIR: an agile and RESTful approach to healthcare information exchange. Paper presented at the Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on. 2013;20–22 June 2013.

Benson T. Why interoperability is hard. In: Benson T, editor. Principles of health interoperability HL7 and SNOMED. 2nd ed. London: Springer; 2012. p. 21–32.

CMMS. Medicare program; revisions to the Medicare Advantage and prescription drug benefit programs: clarification of compensation plans. Interim final rule with comment period. Fed Regist. 2008;73(221):67406–14.

CMS. Center for Medicare and Medicaid Services, Medicare program; identification of backward compatible version of adopted standard for e-prescribing and the Medicare prescription drug program (NCPDP SCRIPT 10.6). Interim final rule with comment period. Fed Regist. 2010;75(126):38026–30.

Cornet R, de Keizer N. Forty years of SNOMED: a literature review. BMC Med Inform Decis Mak. 2008;8(Suppl 1):S2.

Dixon BE, Zafar A, Overhage JM. A framework for evaluating the costs, effort, and value of nationwide health information exchange. J Am Med Inform Assoc. 2010;17(3):295–301.

Do NV, Barnhill R, Heermann-Do KA, Salzman KL, Gimbel RW. The military health system's personal health record pilot with Microsoft HealthVault and Google Health. J Am Med Inform Assoc. 2011;18(2):118–24.

Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL, Essin D, Kimber E, Lincoln T, Mattison JE. The HL7 clinical document architecture. J Am Med Inform Assoc. 2001;8(6):552–69.

Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shabo Shvo A. HL7 clinical document architecture, release 2. J Am Med Inform Assoc. 2006;13(1):30–9.

Ferranti JM, Musser RC, Kawamoto K, Hammond WE. The clinical document architecture and the continuity of care record: a critical analysis. J Am Med Inform Assoc. 2006;13(3):245–52.

Goossen W, Langford LH. Exchanging care records using HL7 V3 care provision messages. J Am Med Inform Assoc. 2014;21(e2):e363–8.

Harding A, Stuart-Buttle C. The development and role of the read codes. J AHIMA. 1998;69(5):34–8.

Health Information Technology for Economic and Clinical Health (HITECH) Act, Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA). 2009.

HL7. FHIR DSTU2: fast healthcare interoperability resources. 2015. Retrieved from http://hl7.org/fhir/index.html.

HL7. Health level seven international: child health. 2012. Retrieved from http://www.hl7.org/Special/committees/pedsdata/index.cfm.

Hripcsak G, Wigertz OB, Clayton PD. Origins of the Arden syntax. Artif Intell Med. 2015.

Huff SM, Rocha RA, McDonald CJ, De Moor GJ, Fiers T, Bidgood WD, Forrey AW, Francis WG, Tracy WR, Leavelle D, Stalling F, Griffin B, Maloney P, Leland D, Charles L, Hutchins K, Baenziger J. Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. J Am Med Inform Assoc. 1998;5(3):276–92.

James AG, Ng E, Shah PS, Informatics Research Committee, T. e. C. N. N. A reference set of SNOMED terms for the structured representation of respiratory disorders of the newborn infant. Stud Health Technol Inform. 2009;150:243–7.

Johnson KB, Lehmann CU, Council on Clinical Information Technology of the American Academy of, P. Electronic prescribing in pediatrics: toward safer and more effective medication management. Pediatrics. 2013;131(4):e1350–6.

Kasthurirathne SN, Mamlin B, Grieve G, Biondich P. Towards standardized patient data exchange: integrating a FHIR based API for the open medical record system. Stud Health Technol Inform. 2015a;216:932.

Kasthurirathne SN, Mamlin B, Kumara H, Grieve G, Biondich P. Enabling better interoperability for healthcare: lessons in developing a standards based application programing interface for electronic medical record systems. J Med Syst. 2015b;39(11):182.

Kaushal R, Kern LM, Barrón Y, Quaresimo J, Abramson EL. Electronic prescribing improves medication safety in community-based office practices. J Gen Intern Med. 2010;25(6):530–6.

Lehmann CU, O'Connor KG, Shorte VA, Johnson TD. Use of electronic health record systems by office-based pediatricians. Pediatrics. 2015;135(1):e7–15.

Liu H, Burkhart Q, Bell DS. Evaluation of the NCPDP structured and codified sig format for e-prescriptions. J Am Med Inform Assoc. 2011;18(5):645–51.

Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc. 2016;0:1–10. (e-pub ahead of print, accessed 3/3/2016 at http://jamia.oxfordjournals.org/content/early/2016/02/16/jamia.ocv189.full).

Nakamura MM, Harper MB, Castro AV, Yu Jr FB, Jha AK. Impact of the meaningful use incentive program on electronic health record adoption by US children's hospitals. J Am Med Inform Assoc. 2015;22(2):390–8.

ONC. DIRECT project. 2014; 7/15/2014. Retrieved from https://www.healthit.gov/policy-researchers-implementers/direct-project.

ONC. Interoperability pledge. 2016. Retrieved from https://www.healthit.gov/commitment.

Reicher JJ, Reicher MA. Implementation of certified EHR, patient portal, and "Direct" messaging technology in a radiology environment enhances communication of radiology results to both referring physicians and patients. J Digit Imaging. 2016;29:337–40.

Rocca MA, Rosenbloom ST, Spooner A, Nordenberg D. Development of a domain model for the pediatric growth charting process by mapping to the HL7 reference information model. AMIA Annu Symp Proc. 2006;2006:1077. See http://www.ncbi.nlm.nih.gov/pubmed/?term=17238696.

Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. A model for evaluating interface terminologies. J Am Med Inform Assoc. 2008;15(1):65–76.

Ruch P, Gobeill J, Lovis C, Geissbühler A. Automatic medical encoding with SNOMED categories. BMC Med Inform Decis Mak. 2008;8 Suppl 1:S6.

Samwald M, Fehre K, de Bruin J, Adlassnig KP. The Arden Syntax standard for clinical decision support: experiences and directions. J Biomed Inform. 2012;45:711–8.

Savage R, Williams W. HL7 version 2.5.1 implementation guide: immunization messaging (Release 1.4). Atlanta. 2014. Retrieved from http://www.cdc.gov/vaccines/programs/iis/technical-guidance/hl7.html.

Shakir AM. HL7 reference information model. More robust and stable standards. Healthc Inform. 1997;14(7):68.

Slee VN. The international classification of diseases: ninth revision (ICD-9). Ann Intern Med. 1978;88(3):424–6.

SMART. App gallery. 2016. Retrieved from https://gallery.smarthealthit.org/.

Sujansky W, Wilson T. DIRECT secure messaging as a common transport layer for reporting structured and unstructured lab results to outpatient providers. J Biomed Inform. 2015;54:191–201.

Wade G, Gotlieb EM, Weigle C, Warren R. Assessing voids in SNOMED CT for pediatric concepts. AMIA Annu Symp Proc. 2008;2008:1164. Similarly, see http://www.ncbi.nlm.nih.gov/pubmed/?term=18998995.

Chapter 4
Data Storage and Access Management

Michal Kouril and Michael Wagner

Abstract One of the most basic challenges arising from the increasing use of electronic data in both clinical practice and research lies in the design and implementation of storage solutions capable of accommodating modern demands. Complex organizational structures, which often cross clinical care and basic research, the often sensitive nature of data, and the ever growing volume of both structured or annotated and unstructured data all motivate innovations, e.g., in identity management, audit trails, monitoring and security, and permissions management. Clinical, translational, and health services research generate very large amounts of data and take place within a complex regulatory environment. Data management, proper placement, including long-term preservation of value, requires careful attention to security, ease of use and access, transparency, and compliance with numerous state and federal laws. It also demands close collaboration and mutual trust between of the IT group responsible for support of research and the IT group responsible for support of clinical care and business operations in a complex research intensive medical center.

Keywords Access control • Audit trails • Data capture • Data management • Encryption • Identity • Security • Data sharing

M. Kouril, Ph.D. (✉)
Departments of Pediatrics and Biomedical Informatics, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Avenue, ML-7024, Cincinnati, OH 45229-3039, USA
e-mail: michal.kouril@cchmc.org

M. Wagner, Ph.D.
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, ML-7024, Cincinnati, OH 45229-3039, USA

Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_4

4.1 Introduction

Many research-intensive medical centers struggle with an ever-increasing deluge of electronic data. While the per-unit cost of storage infrastructure has steadily and rapidly declined, there are significant costs (which can trump the hardware expenses) associated with the technical and managerial challenges of how to keep up with the growing storage demands. Furthermore, the trends of integration of storage with processing increases the complexity even further and puts additional strains on the expertise of the users. Questions are inevitable about the value of data and whether its long-term preservation is justified. At the same time, the increasing volumes of data increase demands on processes to ensure data integrity and security, including the need for audit trails on recording all access requests. Often the local resources are no longer sufficient or developing an expertise to store data on-site is impractical and organizations are turning towards off-site managed resources such as cloud providers. Emergence of cloud providers have profound impact on the industry with far reaching consequences. This chapter relays several of the challenges our Research IT group has faced around the broad topic of data storage and access management, as well as some solutions that have worked well. Much of the content is written purely from the vantage point of a centralized Research IT Services group, i.e., with a focus on a global view of an institution's strategic IT needs rather than from the point of view of an individual investigator or research group. Rather than focus on specific technologies or software solutions, we instead describe the challenges and solutions in general terms. Finally, we write this chapter almost entirely based on our subjective experience and our unique circumstances. Nevertheless we believe that many of our conclusions and strategies described herein are applicable in other settings.

4.2 Classification of Data Types

Before delving into any specific strategies for data storage, we will first describe the types of electronic data typically encountered in (but not specific to) a pediatric medical research institution, their categorization, and how the various types of data can be leveraged to maximize the long-term value of the data. From our vantage point as Research IT Service providers and with our goal of efficient management and long-term data value preservation in mind, we have found it useful to distinguish two broad categories of electronic data, the main criterion being whether and how well the stored data (we will simply call the stored data a file) is described by meta-data.

"Unstructured" Data In traditional file storage systems, an electronic file can only be identified by its file path (the folder it resides in), its name (including, if applicable, its extension), its creation, access or modification date, its size and possibly by permissions. These attributes are typically not sufficiently informative to

allow someone to understand the context and purpose of the data in the file, say, years after it was generated. This includes many cases of plain-text files, but is especially problematic for binary data, which typically require specialized software for processing. Unless the team that originally generated the data is involved and has, say, sufficiently well-kept laboratory notebooks, it could well be impossible to re-process the contents of the file. Unstructured data files clearly present a challenge when it comes to long-term preservation of value, as they are unlikely to be easily interpreted and re-processed long after they were initially generated.

"Structured" Data We call data "structured" if they are described by standardized, consistent, and well-defined metadata (structured information about the data in the file in question). This metadata must contain the minimum necessary information about the data file in order for it to be "useful" in the long term, e.g., amenable to re-processing. Note that the detailed requirements for minimum necessary information differ for each data type, and that much care needs to be taken to define sufficient standards for different data types. The structure of the data and metadata can then be exploited to help, for example, with storage, retrieval, and access. Examples of structured data include well-annotated databases, where specific information is stored in well defined, fixed fields within a record, but also, say, a DNA microarray primary data file that is described by sufficiently rich metadata. Note that many unstructured data files can rather easily be turned into structured data, if processed for example, by a simple parser. Conversely, data that may be amenable to being categorized as structured (i.e., where the required minimum metadata can easily be parsed out directly from the file) may not be identified as such and simply stored on an ordinary network storage device. The latter case represents a missed opportunity of increasing the ease of retrieval and interpretability of this data file and is something to be avoided and/or detected with automated methods.

It is clear that in order to maximize the long-term value of data, an institution should strive to maximize the proportion of structured to unstructured data. This, however, comes with significant challenges, especially in a large pediatric research setting where individual investigators enjoy a great deal of autonomy in terms of deciding on which processes and workflows to implement in their daily operations. The definition and capture of metadata can be burdensome and involve significant changes in processes. End-users who see few or no short-term benefits will likely resist change. Our approach toward overcoming these challenges has been threefold. We encourage the use of user-friendly tools such as electronic laboratory notebooks and web-based data portals that can to some extent automate the generation of metadata (see the following sub-section). Secondly, we made the decision to discourage unstructured data storage by levying charge-backs for networked storage drive, which are not conducive to retaining metadata. Thirdly, we have worked closely with the various institutional research cores that generate large amounts of data (e.g., flow cytometry, DNA sequencing, etc.) in order to ensure that most of the data generated is appropriately categorized, annotated and stored, e.g., in web-based

data portals (see the following section). As a consequence, a large fraction of data can be captured and categorized at the source, and this minimizes the chances of the same data ending up as unstructured files.

4.3 Sample Systems for Structured Data

In this section we will provide examples of ways to capture, store, and share structured data. The requirement that metadata accompany each primary data file makes the use of a (typically relational) database along with a file storage suitable for these data types. We discuss a variety of different options for different types of data in the following subsections. Clearly all require significant investments, whether they are commercially available solutions or home-spun implementations, and there is significant overhead associated with the addition of a database system on top of the file storage solution. However, we believe these are the best ways to maximize the long-term value of data, and we strongly advocate incentivizing their use by investigators.

A database, almost by definition, holds well-structured data, which is decomposed in the case of a relational database into tables, columns and rows. Typically, end-users do not interact with databases directly, instead they usually enter and retrieve data from a database through a middleware software layer. While well-designed databases almost naturally fall into the category of structured data storage systems, care must still be taken to ensure the long-term value of the data, e.g., by ensuring that properly curated data dictionaries are included in the database's documentation. Databases provide many advantages and features to help with fast storage and retrieval of information, such as indexing of values for fast search, optimized underlying data storage, etc. Many database engines also have the capability to control access to the data up to individual tables and rows. To optimize the storage and retrieval mechanisms, especially when working with a large homogenous datasets, databases might not hold the data itself, but merely contain metadata and links to the large data files. Other considerations include the fact that not all data are amenable to being stored in databases, and that there is a significant overhead associated with the *de-novo* design and implementation of a database.

One must also consider the scalability of the database- and, in general, computational-engines – the traditional model of processing relational data is becoming unsustainable for very large datasets or datasets with very high throughputs. A number of frameworks emerged such as Apache Hadoop, Spark, Cassandra, SOLR, etc.

Another factor is the choice to store and process data on-premise or in the cloud. Many cloud providers offer flexibility to choose among many database engines such as Oracle, MySQL, MSQL with a click of a button significantly reducing the installation and setup costs and time. In fact the user can often bring their own license (BYOL) to potentially lower the cost or just migrate the existing on-premise licenses into the cloud.

We have implemented several web-based data portals for data storage for specific data file types, and we briefly outline their functionality in the following subsection. There also are several off-the-shelf commercial or open-source data capture and storage solutions that rely heavily on databases and have found widespread use in medical research centers. A number of these will be discussed in subsequent sections of this chapter.

4.3.1 Web-Based Data Portals

Typical sources of large data streams in a biomedical research institution are the research cores such as flow cytometry, DNA sequencing, RNA expression, imaging and microscopy. These data-generating laboratories are typically centralized facilities used by large numbers of investigators, and they often generate very large numbers of large data files that are inherently amenable to what we term structured data storage. Typical metadata for these primary research data files should include details about the instrumentation that generated the data, details about the experimental setup, the sample sources etc., much of which can be parsed out of the file or extracted from the software that controls the instrument that acquires the data. In an effort to capture these data soon after it is generated, we have opted to develop (separate) web-accessible "data portals" for each data type. These user-friendly, accessible software packages let users upload, store, search for, retrieve and share their primary data securely and efficiently. Advantages include their web accessibility (zero footprint on client computers!), a powerful, flexible search engine which allows searching through a large amount of metadata to retrieve a dataset, the security provided by enterprise authentication (including single sign-on) and user-driven permissions management. Depending on their level of access, end-users can provision others with controlled access to their data. This minimizes the need for sharing data by creating copies on file systems, which represents an unnecessary, wasteful and costly burden on the storage systems. Finally, the web-based portal software can be enhanced with processing pipelines, e.g., to move data to a high-performance computational cluster, thus relieving end-users of the need to tax their desktop computers with memory- and CPU-intensive computations. We have implemented such systems for genotype calling, Genome-Wide Association Studies, microarray processing, and next-generation DNA sequencing. End-users can launch analysis jobs using state-of-the-art processing software on a high-performance computational cluster without requiring in-depth familiarity with Linux-based commands or interactions with cluster batch scheduling systems. Challenges with these systems include the need to help end-users overcome the change in workflow that inevitably comes with the use of any new software system, as well as to convince them to provide and enter sufficiently rich metadata. From an institutional point of view, there is an up-front cost and effort to develop these systems and to define standards for minimal metadata that must accompany any file of a given type. Open-source

solutions like the iRODS system (http://www.irods.org) (Rajasekar et al. 2006, 2009) are beginning to mitigate the need for home-grown solutions.

Many companies are building cloud-based, managed data services to allow customers to take advantage of the hosted infrastructure with a click of a button. One must pay attention to the organizational specifics and carefully evaluate whether potential customizations are possible with reasonable cost. In case of genomic services with increasing (now PBs) size datasets storage and processing is often beyond on-premise capabilities of individual organizations. Such use-cases are then taking full advantage of the elasticity of the cloud expanding the computational resources as needed.

4.3.2 Databases for Clinical and Translational Data Capture

Clinical research data, while quite different in nature than the data types described in the previous subsection, are also very amenable to being stored in systems based on databases. The electronic health record (EHR) can serve as a source of data to support translational research as discussed in Chaps. 6, 8, and 10. To complement the EHR and support research projects by direct data entry, we briefly mention two applications that are used by a large number of investigators in our institution. Many other solutions are available.

- ***REDCap:*** REDCap (Research Electronic Data Capture, http://project-redcap.org/) (Harris et al. 2009) is a widely used, almost self-service secure web-based application designed to support data capture for translational research studies. It is installed locally on web and database servers and provides an intuitive interface for validated data entry, audit trails for tracking data manipulation and export procedures, automated export procedures for seamless data downloads to common statistical packages, and procedures for importing data from external sources. It has been most widely used for studies that do not need to comply with extensive regulatory requirements.
- ***OpenClinica:*** OpenClinica (OpenClinica 2016) is a widely used, web-based open source (LGPL – Lesser General Public License) software for managing clinical trials (https://www.openclinica.com/). It is modular in design and users are permitted to review, modify, and share its underlying source code. Investigators can build studies, design electronic Case Report Forms (eCRFs), and conduct a range of clinical data capture and clinical data management functions. OpenClinica is designed to permit compliance with Good Clinical Practice (GCP) and regulatory guidelines such as 21 CFR Part 11 via differentiated user roles and privileges, password and user authentication security, electronic signatures, SSL encryption, de-identification of Protected Health Information (PHI), and comprehensive auditing to record and monitor access and data changes. There are tools for data cleaning, clinical data management, and site monitoring. Datasets can be extracted in real time and in a variety of formats for analyses.

- In addition to aforementioned free offerings there are a number of commercial vendors (such as Medidata Rave) to provide fully supported environment with appropriate regulatory controls addressed.

4.3.3 Electronic Laboratory Notebooks (ELNs)

Just as software such as Microsoft Office has provided a user-friendly replacement for paper-based documents, one can reasonably expect that electronic laboratory notebooks (ELNs) will eventually replace paper-based laboratory notebooks. ELNs accept input of data from laboratory personnel via computers. Times of data acquisition and changes made in the data can be tracked to meet regulatory requirements. Data is stored in a standard database (e.g., Oracle) and therefore is fully searchable. Modern ELNs will also have ontology-based semantic search features.

In a research environment ELNs are not easily implemented because they require changes to the way researchers traditionally collect and record data. Paper notebooks provide almost unlimited freedom in data collection and in use of free-text. Although the electronic version can be very flexible, it requires a mind-shift, together with the availability of appropriate software tools. Numerous commercial solutions are available.

Our institution chose a particular ELN product in 2007 (CERF by ELN Technologies, Inc.) and has been incentivizing the use of this tool by covering all associated software license fees. Nevertheless, because of the significant changes in workflow required when ELNs are used, adoption rates have been rather low and only a relatively small percentage of eligible investigators have chosen to forgo paper-based solutions.

Cloud based ELN offerings are now equally competitive with the on-premise products. There is an extra complexity related to e.g. regulatory compliance, access security, data storage and movement, etc.

4.3.4 Example Custom Data Capture Systems

It is often the case that no single off-the-shelf software completely satisfies the data capture, storage and processing demands of a particular project. If sufficient funding and local IT expertise are available, then the best option may be to design and implement custom software to support the project. We will describe two examples of complex, grant-supported, research projects that have required extensive resources to plan and develop appropriate data management systems for them.

Genome-Wide Association – Data Management Support Genome-wide association studies (GWAS) use genome-scale information about single-nucleotide variants in large cohorts and correlate the allele frequencies with phenotypic data that is

generally derived from electronic health records. Phenotyping is complex and may require use of Natural Language Processing (NLP) of text in the EHR. Phenotyping is discussed in Chap. 12, Natural Language Processing – Applications in Pediatric Research. The desired outcome of a GWAS study is a list of genomic markers that are (statistically) significantly associated with a phenotype of interest, thus suggesting they be followed up for functional studies. The data required for these kinds of analyses consists of (typically very large) genome-wide Single Nucleotide Polymorphism (SNP) variants assayed on human samples (typically blood or saliva) in conjunction with demographic, clinical and phenotypic data for the individuals in the study. At the present time, SNP data are increasingly being replaced by data from genomic sequencing, which adds to the complexity of data management. The SNP genotypic data typically simply consists of information about whether a sample is homozygous for the major allele, heterozygous or homozygous for the minor allele, which in turn can be easily encoded in a very compressed two-bit format as 00, 01 or 10. The value 11 can thus be reserved for the cases where the data is missing, e.g., for the case where the software that interprets the acquired fluorescence images fails to make a confident call. This leads to a very compressed representation of SNP data and implies that data from chips with millions of genotypes can be stored easily in a relational database as a binary string.

While several excellent packages for GWAS data analysis are readily available (e.g, PLINK, http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al. 2007) and systems such a research patient data warehouse that extracts, transforms, and loads data from the institutions EHR (Data Storage and Access Management, this chapter). Software such as REDCap or OpenClinica can also be used to capture and store the clinical data. It was critical for our purposes to have a system in place that would integrate the clinical phenotypes and the genotypes in one single database and thus facilitate integrated analyses and flexible queries by end-users, who are not necessarily well-trained. Hence, we implemented a web-accessible database using open-source tools such as PHP and MySQL. A relational database stores all demographic and phenotypic information and also holds the genotypes in binary strings. The query interface then permits users to assemble, for example, combined phenotypic and genotypic data for case-control studies, which can be further analyzed using standard open-source software (See Chap. 18, From SNP Genotyping to Improved Pediatric Healthcare).

From the perspective of data storage, which is the primary topic of this chapter, however, it is important to stress that this represents a solution that clearly falls into the "structured data" category: the primary data (genome-scale variant data) is well annotated by individual demographics and clinical data, making it useful for further downstream analysis. If the same data were stored as disjoint flat data files and spreadsheets on a network drive, in all likelihood it would be of little value to researchers in the future who would be unfamiliar with details of data accrual. The web interface and the highly optimized database, while certainly not inexpensive because of the need to repeatedly involve programmers and designers, do provide structure and preserve the value of these data in the longer term.

MRI Imaging Data Management Support A customized data management system was developed to support a specific project that requires integration of neuropsychological test data on normally developing children ages 0–18 years with data from neuroimaging (functional Magnetic Resonance Imaging (fMRI)). The purpose of the study is to establish a normative database of brain development in children that can be used as a reference in studies of diseases. Similar to the GWAS database described above, this study requires storing very disparate data types (clinical, neuropsychological, and imaging) which would be very difficult to link together and ensure their integrity, if they were stored as spreadsheets, text files and images. A relational database, however, together with its web-based interface was developed to provide a uniform, user-friendly gateway for entering, storing and processing all data pertaining to the study. As with the GWAS database, role-access is provisioned via an enterprise identity management system, all data changes are logged. In contrast to the GWAS database, data can be entered directly by study coordinators through the web gateway, using double-data entry where appropriate (if transcribing from paper-based forms). fMRI data is even larger and more unwieldy than genome-wide genotyping data, and so in this case the data are stored in protected files on the file system and only references to the files are stored in the database. The web interface supports a very general search function that permits users to query the more than 500 variables stored in the data and to retrieve images satisfying search criteria. The data can be downloaded and processed offline, or it can be submitted directly from the web interface to the LONI Pipeline (http://pipeline.loni.ucla.edu/) (Rex et al. 2003) workflows, which run on the local Linux cluster. For more information the interested reader is directed to the project web site: http://research.cchmc.org/c-mind.

To summarize this section, both of the examples provided are instances of large, well-funded projects that warranted custom data storage and management solutions. In both cases the databases along with the respective web interface facilitate access to diverse data types and very large, unwieldy files in a central repository. The ability of end-users to directly access, search (using rich meta-data!), download, and process complex data sets, the built-in security, audit and quality control features as well as the ability to access the data easily with only a web browser have made these tools indispensable. The initial design and development costs are certainly high, but these can be amortized by offshoot projects that only require relatively minimal modifications to the initial designs (e.g., other fMRI research studies).

4.4 Unstructured Data

As examples of unstructured data storage options, we briefly describe network attached shared (NAS) disk data storage and the newly emerging online cloud storage technology.

4.4.1 Personal and Shared Network Storage

Personal and shared network storage referred to as NAS (Network Attached Storage) is generally set up to mimic a local hard drive. Similarly to local drives, NAS solutions allow users to store flat files without imposing any requirements on metadata or other structure. NAS servers typically provide added benefits including sharing stored files and data with multiple people or devices via a network. NAS devices are in general centrally managed and backed up by an IT group, thus eliminating the need to manage backups on individual workstations. Often NAS devices also support audit functions, which means that a history of all operations on the file system, including who performed them and from where, is stored. This facilitates (but is certainly not sufficient for) compliance with regulations such as HIPAA (HIPAA 1996). The resiliency of the unit is often favorable compared to local storage. NAS units are often able to withstand the loss of one or more disks before an actual data loss occurs. To provide physical security, NAS servers are typically located in data centers, which protect them from theft and unauthorized physical tampering.

There are two leading protocols to access data over a network: CIFS/SMB (Common Internet File System/Server Message Block) – often referred to as Windows based file sharing – developed by Microsoft and NFS (Network File System) – Unix/Linux-based file sharing – originally developed by Sun Microsystems.

Network transfer speeds are improving and the times to store and retrieve data from NAS systems are becoming comparable to a local hard drive. Potential disadvantages of using NAS devices are related to the complexity of the NAS setup (which requires some in-depth technical expertise) and their maintenance (compared to local disk). Moreover, NAS systems by default generally do not support tying metadata to the primary files, and thus the associated ILM (information lifecycle management), if one exists at all, is somewhat less effective than ILM implemented over structured data. NAS storage is currently the most common mechanism to store data in medical research centers. Personal network data storage is typically restricted to files that do not need to be shared. Shared network data storage on the other hand typically contains business and research files that do need to be shared.

4.4.2 External Hard Drives and USB Flash Drives

External hard drives (sometime called USB hard drives) are popular because of their low price and portability. Their use is strongly discouraged within an academic health center, except for special purposes such as encrypted data transport. Their major drawbacks include:

- General lack of support for encryption
- Poor physical security (easy to steal)
- Prone to failure and typically do not support any kind of internal redundancy
- Can only be connected to a single computer

Flash drives (sometime called USB memory sticks) are often used for data transfers, presentations, and movement of data sets from one computer to another or possibly to another institution. This flexibility has to be accompanied by strong encryption because the stored datasets in an academic health center often contain Protected Health Information and flash drives are very easy to lose. In order to use USB flash drives properly, an institution should strongly consider deployment of an enterprise encryption solution for USB flash drives.

4.4.3 Online and Cloud Storage for Unstructured Data

To address the ubiquitous needs for sharing data and easy availability and accessibility of data anytime and anywhere, several companies offer so-called cloud storage. In reality these cloud-based storage solutions are (merely) internet-accessible servers with large amounts of storage. Cloud storage vendors often develop clients for multiple platforms so that a user can access the same file from multiple sites, e.g. a laptop at home, a workstation at work, a smart phone, or a tablet computer. The major advantage of using online or cloud storage is the users' ability to access stored data from virtually anywhere and to share data with other users.

There are concerns for an institution adopting an external cloud solution compared to deploying an in-house solution. First there is an issue of compliance with laws and rules governing access to and use of many types of information, such protected health information. Institutions and vendors must assure compliance before allowing general use of cloud storage. Second, transfer speeds can be a problem with bottlenecks, possibly on a vendor side or related to connectivity between the vendor and the institution. Problems arise with transfer of the huge data sets characteristic of many types of biomedical research. Last, but not least, there is the question of costs, including subscription, internet bandwidth, support, etc.

4.5 Compliance and Retention

With any type of data storage (structured or unstructured), retention, compliance, and security are always of concern, both for investigators and for host institutions. Strategies to deal with these issues differ for different types of data. Our institution and projects are often a subject to HIPAA, FISMA, 21CFR-Part11 compliance, various state- and project-specific data use agreements, etc.

4.5.1 Disaster Recovery

Strategies for disaster recovery for different data types can take many different paths based on the criticality, sensitivity, and amount of data that must be protected. Traditional methods for backing up and recovering data (such as tape drives) may be adequate for smaller data sets, however these are becoming increasingly difficult to utilize because of the very large size and scope of datasets that are commonly generated in translational research. A number of key considerations must be taken into account before establishing a strategy that is appropriate for the environment.

- **Business Impact Assessment (BIA):** The BIA will document the criticality, impact, and sensitivity of the data in relation to other applications, systems, and data within the environment.
- **RPO vs. RTO**: The BIA planning usually operates with a couple of metrics which differ among an application based on the business impact of how much data does the business withstand to loose during a disaster as well as how quickly must a particular application be up and operation. The first is called Recovery Point Objective (RPO) and indicates how recent must the data be in the restored application (1 h RPO means that the institution can only loose last hour of data prior to a disaster). The second metric is called Recovery Time Objective (RTO) and indicates how quickly we must restore the service (1 h RTO means that the institution expects the application to be up within 1 h disaster). Usually the shorter the RTO or RPO the more complex and expensive technology is used to protect the data and applications.
- **Types of recoveries:** The types of recoveries that are going to be required must be cataloged. Strategies for recovery of structured and unstructured datasets can be different. With structured data, the recoveries in all likelihood involve backing up and restoring entire databases, which can imply that very large amounts of data need to be copied on a regular basis, and maintained in the environment. With unstructured data, the files that are generated and need to be restored in the event of a recovery may be small in size, but could be large in number.
- **Implementation:** Once the BIA and types of recoveries are established, a solution needs to be implemented. Based on the criticality and impact of the data, this solution will typically be multi-faceted and have several guises. A disk-to-disk and disk-to-tape solution will work for most situations. In the disk-to-disk scenario, backups are performed using secondary disks to store the backups for rapid recovery and speed of backups. For longer-term retention and where data volume allows it, the backup set on the disk is transferred to tape for off-site storage and data retention.

4.5.2 Retention

Data retention strategies are among the most difficult to implement in the context of institutional data. Most data is generated and then stored indiscriminately because of complexities related to establishing institutional retention policies or strategies over all data generated by an institution. This leads to storage of large amounts of data that are used once or never used. This is an expensive waste of institutional resources. When establishing an institutional policy for data retention, it is important to keep the following key considerations in mind.

- **Usability:** How long does the data maintain its usability for the institution? Many data sets, once generated and analyzed, can be destroyed because they are no longer useful to investigators. On the other hand, some data must be retained to meet regulatory requirements or to document critical research or business reports. Institutions should establish policies with time frames for retention of data derived from different sources for different purposes.
- **Impact:** In the event of a breach or loss of data, what is the potential impact on the institution? If data is no longer needed, then it should be destroyed so that it is no longer a potential source of institutional liability should a breach occur.
- **Privacy/Compliance:** Protected health information (PHI) is typically stored to support translational research. Regulations require that once such data is no longer needed or if the owner of the PHI requests its removal, then the data must be deleted. Data may be subject to other types of regulations, such as those promulgated by the Food and Drug Administration for data related to clinical trials or by the Defense Department or Homeland Security.

Once an institutional policy and a strategy are developed, they must be implemented. Processes must be in place to allow for the removal of data elements, once they are no longer needed or have reached their retention lifetime. The content of files containing unstructured data is generally poorly documented. This poses major challenges in file management. Typically, the files containing unstructured data will need to be indexed before decisions about deletion can be made. For structured data, indexes exist and can guide the removal of data elements from tables/structures.

4.5.3 Encryption

Encrypting structured and unstructured data can be difficult, depending on the format of the data, how it is accessed, and how it is transported. File access protocols such as CIFS do not have encryption built-in, which can make it difficult to protect the data when being transmitted. To make it more difficult, compliance standards may require that protected information be encrypted when being transported or stored.

- **Transport (data in flight):** Encrypting the data while it is transmitted across networks and between systems.
  - This provides protection of data during transport so that malicious users cannot intercept and read it
- **Storage (data at rest):** Encrypting the data when written to systems for storage and access.
  - This provides protection against lost disks or other physical attacks against storage locations

For unstructured data, depending on how the data is accessed (CIFS, NFS, HTTP), services can be added to provide encryption. When working with protocols such as CIFS, encryption is not inherently built-in so additional tools such as Virtual Private Networking (VPN) must be implemented to provide the encryption. By moving the storage location behind a VPN device, all access to unstructured files using the CIFS protocol will be transmitted down an encrypted tunnel to the CIFS volume thus providing the level of protection required. Other protocols such as NFSv4 and http allow the addition of encryption to the protocol, using built-in standards (e.g. HTTPS, NFSv4 encryption). When working with structured data, access is generally by protocols that support encryption (HTTP(S)). Since the application is presenting the data, encryption such as SSL/TLS can be added to the protocol to encrypt the data in transport from the application.

Encrypting data storage can be achieved through different methods and must be applied to all locations (e.g., primary and backup locations). Tools such as Full Disk Encryption (FDE) and Full Database Encryption (FDbE) can provide the encryption, when written to systems. No matter what method of encryption is used, there are important items to remember when implementing encryption in an environment:

- **Key Management:** When implementing encryption, it can only be as secure as the keys that are used to encrypt the data. Use of weak keys or poor protection of keys, can allow attackers to decrypt data and bypass the encryption. It is essential to choose strong keys when encrypting data and ensure that key access is limited to those individuals who require access.
- **Key Rotation:** Encryption keys, like any other password, are susceptible to compromise. While keys may be large enough to make this difficult, with computing resources available, they may be guessed at some point. To address this, keys should be rotated at least periodically to reduce the possibility of their being guessed and used to decrypt protected information.
- **Performance:** Encryption requires additional processors to handle the encryption/decryption operations. In high-volume environments, encryption places significant overhead on servers, reducing the overall performance of the system. Where possible, hardware devices such as Hardware Security Modules (HSM) and SSL Offload should be added to the system to handle the encryption/decryp-

tion, since they have cryptographic processors and are purpose built to handle these operations.

4.6 Management of Permissions and Access

Translational research generates huge amounts of data that is sensitive and must be access-protected. While there are extensive regulatory requirements regarding the protection of data from human subjects, it is clear that other types of data must be stored securely for a variety of reasons, for example, to protect investigators' intellectual property rights or to meet requirements of state and federal agencies that support research in the United States, such as the NIH, CDC, FDA and the Departments of Defense and Homeland Security. Furthermore, investigators must be able to ensure the integrity and validity of their electronic data, which would be impossible without tight control of access to their data.

A less obvious question is that of data ownership and responsibility for data protection. While host institutions usually hold claim to all data generated by their employees, academic investigators typically retain a great deal of independence and demand that they control access to all data generated under their supervision. This is typically very different from a corporate environment, where the company retains complete control and ownership of all data and thus can decide how access to sensitive data will be provisioned and justified.

4.6.1 Permissions

The complexity of large research organizations fundamentally dictates that any data storage solution (be it for structured or unstructured data) have a well-defined process for permissions and access management. Such a process is typically anchored in an institutional identity management solution, which uniquely identifies the data every defined entity (individual, application etc.) is permitted to access. Ideally, a self-service solution is in place that allows end-users to request access to data and resources. Changes in permission for an individual or group become effective as soon as all required authorizations have been granted, that is, the process between the request for access and the authorization and implementation of change is smooth and timely. Changes in permission to access should be audited and be transparent to both end-users and data owners. In practice, the process to grant permissions is generally better defined than the process to deny or withdraw permission, as when a user leaves the institution or moves to another department. This issue can be mitigated by carefully tracking the owner of data (e.g., a protected folder, database, etc.) and requiring periodic entitlement reviews to confirm the current list of provisioned users.

A specific example of a self-service permissions management system for unstructured data storage that works well in the author's institution is the commercially available DataPrivilege package (Varonis Inc). The initial setup involves defining all protected directories on the file systems and importing them into the permissions management system. For each protected directory, a list of authorizers must be defined. They have the power to confirm or deny requests for permission to access data. Once the system is setup, users can login to a webpage to request permissions, which triggers an approval workflow potentially resulting in provisioning access. This workflow does not require involvement of IT specialists and is entirely self-service. An IT team is required only to keep the system running, monitor for unauthorized changes, and help users setup new protected directories. This system has worked especially well in our setting because investigators can fully control and oversee permissions for all folders containing their data. Furthermore, the audit trails in the system, as well as the various monitoring functions, allow staff to find rare cases of unauthorized changes in permission.

Standard database engines (such as those used for web-based storage portals and clinical research data capture system) offer role-based access for the management of permissions. If *ad-hoc* applications provide the front-ends to databases, then compliance with institutional guidelines for control of permission and access should be followed.

4.6.2 Access Auditing

In addition to processes for data protection and access, it is important to have an audit infrastructure in place to capture who is accessing data, when they are accessing it, and how they are accessing it. Audit data is important when investigating unauthorized changes in data, security breaches, and failures of compliance. Audit capabilities need to be enabled within the applications and systems hosting the structured and unstructured data. This can be as easy as enabling built-in audit features or can require additional development within applications to document access to data hosted by the application. In either case, once the audit data is enabled or generated, the audit data needs to include the following items to ensure it can be used. The fields are built-into most protocols such as the Syslog and Windows event log:

- **Date/time of event:** All audit data should include the date/time of the event. This should be synchronized with other systems to ensure consistent time across events within the environment.
- **Log source:** The location/facility that generated the event (e.g. server, application, etc).
- **Outcome:** Whether the event was successful/failure.
- **User/system:** The users and/or system that generated the event.

Once the audit data are generated and stored it is beneficial to correlate it with other events within the environment. Tools such as Security Incident and Event Management (SIEM), Security Incident Management (SIM), or Security Event Management (SEM) can be used to collect, store and correlate event data from multiple sources. These tools will take the event data and normalize it with other sources in the environment to provide a consistent picture of the event data and can be used to alert personnel in the event of pre-defined rule violations. Audit data are typically reviewed on a regular basis so that suspicious events are not missed and possible breaches are quickly detected. Event data should be maintained within the event correlation systems for at least a month and can then be rotated to offline storage. This permits timely investigation of unauthorized access. Offline storage should be maintained for at least a year to ensure records can be restored and reviewed when required. This will provide adequate time to perform investigations and to meet compliance requirements.

4.6.3 Sharing Data with External Collaborators

While investigators need convenient options for sharing information with colleagues, host institutions need to protect themselves and their staff from data leakage, unintentional exposure of often-sensitive data, and hacking. Data can be transmitted by email or uploaded/downloaded using many of the data transfer protocols such as FTP and HTTPS. Content management systems such as Microsoft SharePoint, or the open source Drupal and Plone content management systems (CMS) have capabilities to securely share data files, but are not often optimized for large data sets. Xythos (Blackboard, Inc.) is an example of a commercially available application to facilitate data sharing. It provides a web interface to local storage with many convenient features, such as quota management, data sharing via web tickets, automated work flows, subscription services, and automatic expiration of access provisions.

E-mail, despite being a technology that is now several decades old, is still one of the most common ways to share data. Email is transmitted as plain text unless attachments are explicitly encrypted. Email servers have built-in size limitation, which is often the reason why researchers switch from email to an online storage for file sharing. Most institutions employ so called spam filters, which remove unwanted or harmful emails. Such filters can interfere with an exchange of attachments. There are commercially available encryption devices, which ensure that the emails and attachments are delivered to the intended user over an encrypted channel. In order to address limitations inherent to email, alternatives such as ZIX (http://www.zixcorp.com/) or Xythos are often deployed on premises to complement email. Besides the on-premise alternative, cloud solutions, such as Microsoft OneDrive, DropBox and Google Drive, can be used for certain types of data, provided that the institution enforce appropriate conditions and limitations.

Data availability itself is often not enough and we end up looking for the right solution. HTTP/S protocol is very flexible, but not very efficient for large datasets being transferred over medium to large distances. Application integration with high speed application transfer software such as Aspera or Data Expedition Expedat is necessary to be successful in sharing large data sets.

In addition to availability and speed we have seen many use-cases for managed data transfer orchestration. During the transfer the service is capable of performing additional steps on the transferred piece data to improve security, compliance and ease of collaboration.

4.7 Managing Growth

Translational research programs pose enormous challenges in data storage and management because of the enormous growth in the quantity of data that is being generated. For example, because next-generation DNA sequencers and confocal imaging microscopes are capable of generating terabytes of data in a matters of days, the focus is shifting from how to create petabyte size storage servers (a relatively simple task) to deciding how to effectively *manage* what is stored on these servers and how it is stored. The key is exploring the data itself – what do they contain and can we leverage metadata to create an effective strategy for information lifecycle management.

4.7.1 Managing Growth of Structured Versus Unstructured Data

Structured data, by definition, is annotated by metadata and thus can be incorporated more easily into an information lifecycle. The implication is that the decisions around what to require for specific data types in terms of their associated metadata are crucial and have a high impact on the ability to manage the data effectively. One should attempt to find a balance among the quantity of information that needs to be stored, how it should be stored, how it will be backed up with appropriate methods of disaster recovery, and associated costs. Data-generating research cores such as the flow cytometry facility, confocal imaging core, sequencing cores and others represent a large fraction of the demand for storage. Therefore, it is especially desirable to ensure that all data generated by these cores can be stored in well-structured and annotated forms, making it conducive to cost effective strategies for information lifecycle management. Typically, the growth in demand for storage by these cores is fairly predictable, using simple assumptions such as average core utilization rates.

Of greater concern is growth in unstructured data, which can arise when users download (often large) data sets from public sources without annotating them or store duplicate copies of basic research data, which also resides in a repository in structured form. Growth in unstructured data is often unpredictable, especially if data storage is made readily available, and users have little or no incentive to be selective in what they store. The challenge for IT service groups is to steer growth from storage of unstructured data to storage of structured data with accompanying metadata.

4.7.2 Chargebacks and Quotas

Chargebacks create a financial incentive to minimize data storage and to delete data when it is not longer needed. For example, research groups must decide whether to keep all intermediate data related to an experiment or only those that can not be recreated easily. The investigator must weigh the costs of storage (through chargebacks) against the cost to re-generate the data. Users can be assigned disk space they "own", i.e users pay for allocation of a disk, which they then manage, rather than paying for storage as needed, which is assigned and managed institutionally.

A storage quota typically refers to the maximum amount of data a user can store. Once the user crosses the pre-set threshold, the system can lock the directory, file system, etc. and prevent the user from writing additional files. Alternatively, crossing the threshold can trigger an alert to notify the users, who then negotiate either a quota increase at increased charge or file removal.

4.8 Joint Management – Hospital and Research

Translational research typically occurs in a clinical environment. IT services in support of research may be managed by a hospital or by a separate academic organization. Hospitals typically have a large IT staff to support clinical care and business operations, especially if they have implemented an electronic health record. In theory, the hospital IT group could support the less demanding needs of research programs. In practice, this poses challenges because support of clinical care and critical business functions will always be of higher priority than support of research. In addition, research programs generally use many different customized software applications and platforms, while hospital operations typically are supported by products of commercial vendors. To support research in a clinical environment, it is important to define clearly who is responsible for providing IT support to the different missions – patient care, business operations, and research. In the authors' institution, clinical and business operations are supported by hospital information services (IS), while research is supported by biomedical informatics (BMI). The two groups are separately staffed and budgeted, but must collaborate very closely to

Table 4.1 Distribution of responsibilities for support of research between Hospital Information Services (IS) and Biomedical Informatics (BMI)

Issue	IS	BMI
Centralized enterprise data storage for research	X/J	P
Specialized data storage for research	J	X/P
Backups and archives of research data	X/J	P
Security	X/P	J
Authentication	X/J	P
Research applications hosting	J	X
Research web server hosting	J	X
Open source and non-enterprise software for research	J	X
Support for research applications	J	X
Support for enterprise applications that support research administration (e.g. grants administration, Institutional Review Board, animal care)	X/P	J
Content ownership, end-user training & support for enterprise research applications	J	X/P
Software evaluation and development for research	J	X/P
Clinical software implementation	X/P	J
Enterprise software evaluation & implementation to support faculty research (e.g. Electronic Laboratory Notebook, REDCap)	J	X/P
Manage licenses for commercial enterprise software	X/P	J
Manage licenses for research non-enterprise software	J	X/P
Help desk 24/7, Desktop support	X	J

X primary responsibility for operational support, *P* primary responsibility to formulate policies, *J* participates in formulation of policy

provide coherent support to the overall research and clinical enterprise. The two have developed a responsibility matrix (Table 4.1), which has served the organization well and prevented misunderstandings. A similar matrix of responsibilities should be developed by institutions conducting translational research that requires collaboration among components of the organization or with other organizations (such as a medical school, university, hospital, clinical practice group, research foundation). In our case, the research group has taken advantage of the hospital's need to have a secure 24/7/365 data center and has incorporated a research data center in the same physical facility. Research has service agreements with the hospital to take advantage of the hospital's facilities and staff for data backups and disaster recovery. Because translational research requires extraction of data from clinical systems (see for example, Chaps. 6 – Data Governance and Strategies for Data Integration and 10. Informatics to Support Learning Networks and Distributed Research Networks), BMI complies with the hospital's policies and procedures related to security and protection of information (Table 4.1).

Acknowledgement This publication was supported in part by an Institutional Clinical and Translational Science Award, NIH/NCATS Grant Number 8UL1TR000077-04. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377.

HIPAA. Health insurance portability and accountability act of 1996. Washington, DC: U.S. G.P.O.; 1996.

OpenClinica. "OpenClinica." 2016. Retrieved March 21, 2016, from http://www.openclinica.com.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559.

Rajasekar A, Wan M, Moore R, Schroeder W. A prototype rule-based distributed data management system. HPDC workshop on "Next Generation Distributed Data Management". Paris; 2006.

Rajasekar A, Wan M, Moore R.. Event processing in policy oriented data grids. Proceedings of intelligent event processing AAAI spring symposium. Stanford; 2009. p. 61–6.

Rex DE, Ma JQ, Toga AW. The LONI pipeline processing environment. Neuroimage. 2003;19(3):1033.

Chapter 5
Institutional Cybersecurity in a Clinical Research Setting

Michal Kouril and John Zimmerly

Abstract The principal challenge facing IT groups that support research on a daily basis lies in striking a fine balance: On one hand researchers must share data and use cutting edge analytic tools on a variety of computing platforms to enhance their creativity and productivity. On the other hand, much of the data that supports translational research contains personal health information derived from patients' medical records. Hospitals are justifiably concerned about the highly sensitive nature of the data and the need to comply with a myriad of federal, state, and local laws, and contractual regulatory requirements that demand high levels of security and access control. A number of frameworks exist to help with the process. In this chapter we discuss these challenges and the approaches taken at a research intensive children's hospital to put a policy framework in place that enacts standards for security, risk evaluation and mitigation, monitoring, testing, and design of the IT infrastructure. These protect the institution, while enabling collaboration and innovation among researchers. We stress the organizational need for a close and collaborative relationship between IT groups that support research and those charged with support of the medical center's clinical and business operations. It is also important to recognize that technology alone cannot assure security. Institutional policies and user security awareness education also play key roles in assuring that confidential information is in fact protected.

Keywords Access control • Authentication • Cybersecurity • Data-centric • Data sharing • Firewall • Network • Operating systems • Protected health information • Security • Security awareness

M. Kouril, Ph.D. (✉)
Departments of Pediatrics and Biomedical Informatics, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Avenue, ML-7024, Cincinnati, OH 45229-3039, USA
e-mail: michal.kouril@cchmc.org

J. Zimmerly
Division of Information Services, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, ML-9009, Cincinnati, OH 45229-3039, USA
e-mail: john.zimmerly@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_5

5.1 Introduction

Translational research typically requires access to data that reside in geographically distributed data warehouses that are called upon by a team of collaborators using a variety of software applications. Information technology support of this type of translational research requires networks to permit data sharing, while maintaining high levels of security. Without networking, investigators cannot access the applications, servers, and other resources in a distributed environment. Given this criticality to providing services, networking can be the major infrastructure that provides an environment where access to data is restricted to authorized users and the overall system is protected from malicious attacks by unauthorized individuals (data breaches). According to the Verizon 2011 Data Breach Report (Verizon 2011), which outlines the incidents investigated by Verizon security response teams and publicly accessible data, the top 2 of 10 threats to systems involved direct hacking against servers. Given this threat, network security and sound network security practices can provide a large layer of protection against current threats to the security of the environment.

5.2 Secure Network Design

In this section, secure network design practices along with models for architecting secure networks will be covered including data-centric network, firewalls, intrusion prevention and detection systems (IPS, IDS), and secure remote access including virtual private networking (VPN).

5.2.1 Data-Centric Networking

The traditional model for building enterprise networks takes into account the different trust zones within which applications must be accessed and published. These typically include: (1) the internal zone for trusted employees and machines on the Local Area Network (LAN), and those authorized to remotely access the network; (2) the Demilitarized zone (DMZ), a subnet that provides limited connectivity to the internal network; and (3) the external untrusted Internet zone. This model has worked well for most organizations, providing a barrier between internal and external environments, as well as segregating different populations of users. However, access requirements for most applications are evolving, as technology changes so this model may not work for all situations. Medical centers are finding that users are increasingly using mobile devices, requesting access from all locations, and requiring more help with troubleshooting, when accessing applications from different locations. Given this, the barriers between the internal and external environment are

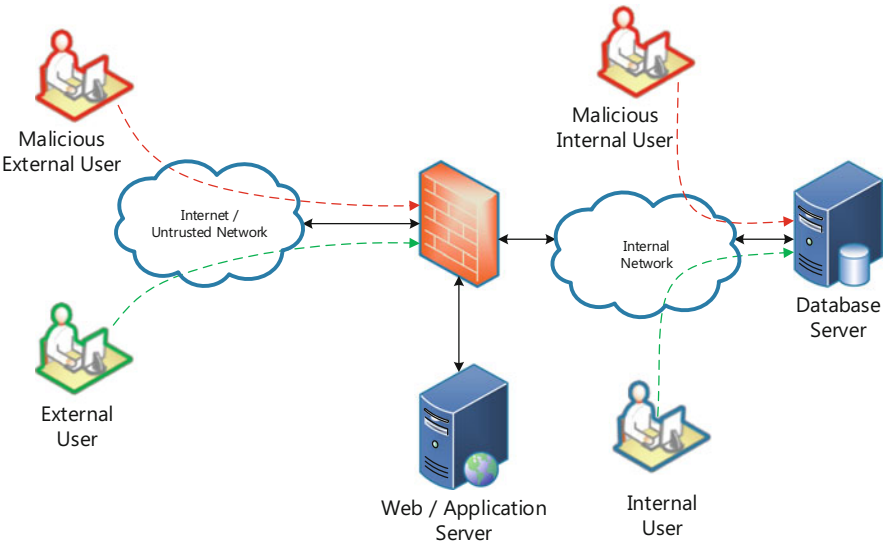


Fig. 5.1 Typical architecture of a network with different zones of trust

becoming less clear and less stringent than they were originally intended. To address this, the concept of data-centric networks has evolved. The aim is to re-architect networks around protecting data, unify the internal/external experience, and build core protections for what must be protected; the data.

Figure 5.1 outlines the setup of a typical network with the different trust zones, using an application server that has a database backend. External users and any malicious users on the Internet will be restricted from accessing the database server directly and have limited access to the Web/Application server through firewall filtering. However, any user who is on the internal network (malicious or not) will have direct access to the database server, where the application data may be stored. Given vulnerabilities or misconfigurations that may be present within the server or database application, users may be able to bypass the application protections and pull data directly from the database, whether they have been authorized or not. Given the scale and size of most enterprise networks, the malicious user in Fig. 5.1 could be accessing via remote offices, affiliates, or other locations that may not have the same level of physical security as a main campus or office may have.

To address the risks posed by the direct access to data as shown in existing network design, data-centric networking looks at building rings of protection around data. Figure 5.2 provides a high-level overview of this design philosophy. This architecture follow similar security designs used within CPU privilege rings (so-called Ring 0 access) (Cruse 2016). Starting from the center of the circle and moving outward, the rings and access become progressively less trusted. No user or process can jump a ring without going through a ring above it. This forces all access

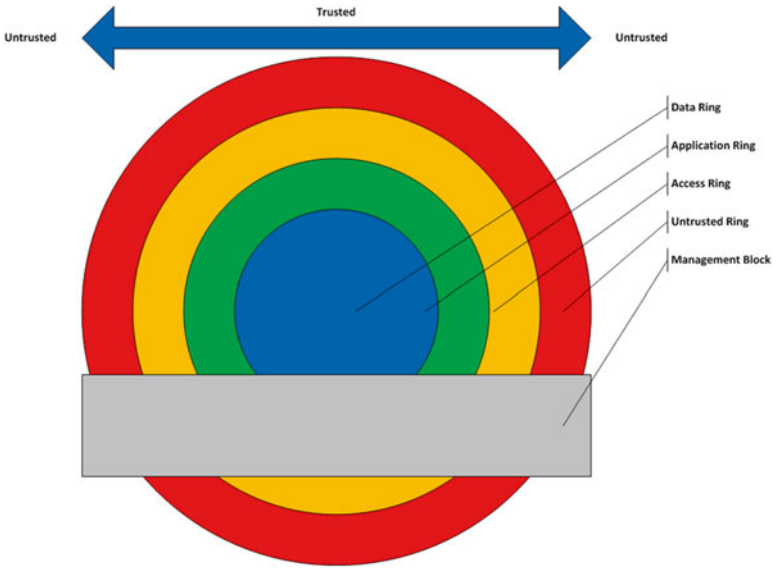


Fig. 5.2 Building rings of protection around data

to data through known applications and paths that can be hardened, monitored, and protected using internal controls. This layering approach can also help in the principal of containment of malicious activity. Rings include:

- Data Ring: This would include all structured and unstructured data sources such as database servers, file servers, etc.
- Application Ring: This would include all application servers serving content or publishing data that is accessed, used, or manipulated by users.
- Access Ring: This is the main ring that presents the initial login, authentication, and authorization to the users. This would typically include VPN, proxy servers, etc. that are used for access to the environment.
- Untrusted Ring: This is anything outside of the environment. This can include both the Internet and internal network segments.
- Management Block: This would include management systems used by system administrators to maintain the environment, such as patching, authentication, backups, etc.

Figure 5.3 shows a network with the same type of users as the network in Fig. 5.1, but has been re-architected in accord with the data-centric network model. As shown in the Figure, there is no direct access to the database server or application except from the management layers. Users can access an application only through the access server, which will present the data back to them. All access to the application is through the external firewall, to the access server, and then to the application. At these initial rings, filtering for users, locations, etc. can be put in place to further limit the scope of access by users. None of the traffic is ever sent to the application

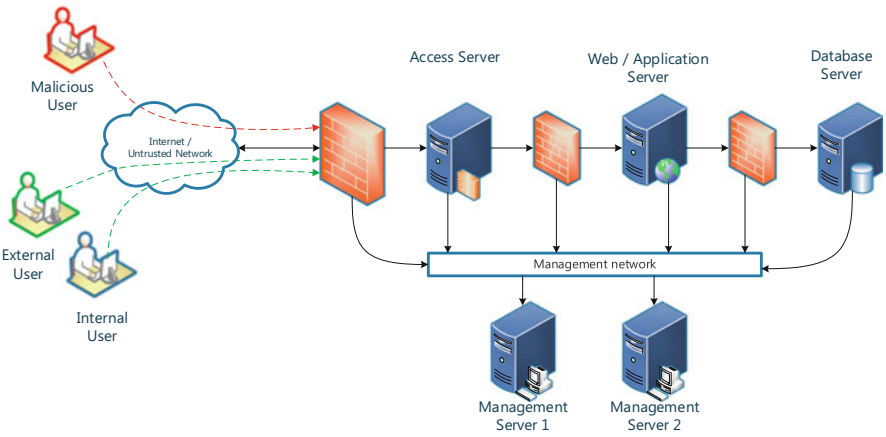


Fig. 5.3 The network shown in Figure 5.1 has been re-architected in accord with principles of the data-centric network model

or database server until it has been properly authorized to be legitimate traffic within the environment. The data-centric network model reduces the access footprint on the database server from the typical 50–200 ports that may be accessible in the traditional model down to zero ports because only the application server can access the database server. This reduces the risk of unauthorized internal users accessing the database server. It also decreases the possibility that malicious users can gain access to the internal network or access the server to launch attacks. Some of the other benefits of this model include:

- Define Firewall Access: Since all application and data access must go through a firewall interface, specific inventories of port usage are gathered during the deployment of applications and can be used for auditing and compliance, when reviewing servers and applications in the environment.
- Unified User Experience: A single method of access is used by all authorized users regardless of location, i.e. the process of access is the same, whether they are at an internal (e.g. at work) or external (e.g. at home, travelling).
- Enforced Standards/Monitoring: Since all access to the applications from within the internal and external zones is funneled through a limited number of access methods in the access ring, more monitoring of access and enforcement of standards can be enabled to reduce the potential of successful attacks on the environment.
- Outbound Access Protections: Since all the application and database servers are on internal limited access segments, outbound access from the servers can be denied. This will reduce the possibility of an attacker gaining access to the system and installing software to copy data to outside sources. A user who gains access to the internal network through the firewall can only access services/resources that have been pre-authorized. This reduces the possibility of data escaping the environment.

- Reduced Horizontal Attack Surface: Since most environments are only as secure as their weakest server or application, limiting access between rings and systems within the rings reduces the possibility of a lower-security application server being compromised and then used as a pivot point to launch attacks against more secure systems.

There are a number of challenges that must be overcome in using the data-centric network model. These include:

- Performance: In a typical application/database scenario, data access is relatively simple and is over a switched network or routed across a core router (Fig. 5.1). Typically, performance is excellent. In a data-centric network, all application database access must be sent through a firewall with limited access (Fig. 5.3). This can impact database access times and cause application performance issues based on the bandwidth the firewall can handle. Large infrastructure firewalls or firewall service modules in core routers/switches can address this and allow for setup of different network and trust zones.
- Network Re-Architecture/Change: Implementing the data-centric model requires a re-design of existing environments and for example, moving servers, retraining staff, rerouting connections, and implementing additional firewalls. This is expensive,can be quite stressful to the IT staff, and can impact operations
- User Access: When database servers are on internal networks, users may become complacent/accustomed to accessing servers directly when they are in their office/workspace. In the new model, users at the office, as well as at home, will be required to VPN or use other remote access methods to access databases and other unstructured data within the environment. Users may require substantial training to learn new processes.
- Firewall Changes: When there are no firewalls or other protections in place, server owners and administrators may not have accurate inventories of what ports are required for an application to run. When moving to the data-centric model, issues can arise because non-standard ports are being used and may not have been added to new firewall rules, thus causing issues with application access/operation. Proper sniffing and network monitoring can help identify the ports so that firewall policies can be established before applications are migrated to the new environment.
- Application Development: Developers writing applications for the data-centric environment need to make sure they are writing their applications to use standard database ports or confined port ranges to ensure they can work through different layers of firewalls. When ephemeral ports and other dynamic ranges are used, this can cause issues in negotiating firewall policies.
- Outbound Application Access: Many applications today are built to have auto-update or other web service integration to pull in and process data used within the application. In an environment with limited to no outbound-access, application calls to external resources must be inventoried and allowed to use the infrastructure or other proxies to access. Open source products such as Squid Proxy

(Squid 2016) or built-in proxies in firewalls can be used to allow access to a restricted number of external resources.
- Desktop applications: Many application are design to run on user desktops with connectivity to a remote database or filesystem. In the data-centric network model these application are usually moved to the VDI (Virtual Desktop Infrastructure) setup residing within the protected network. Alternatively a VPN connection is granted to the backend data servers for those users. Generally we try to avoid granting a direct access through the firewalls to the backend resources – if it becomes necessary very narrow firewall rules are highly recommended.

When implementing the data-centric network model, it is important to assess the environment and use of applications within the environment. Given the right architecture and use of the model, the footprint of threats to applications and data within the environment can be drastically reduced with improvement in security.

5.2.2 Intrusion Prevention/Detection

Once the secure infrastructure and firewalls are properly implemented, unwanted traffic is stopped. Given the frequency of attacks and the number of potential vulnerabilities in all systems, best practices dictate that monitoring be in place to alert when possible attacks are taking place within the environment. Continuous monitoring will help respond to attacks in a timely manner, reduce the potential impact of the attack, and reduce the possibility of the attack occurring again within the environment. To achieve this, Intrusion Prevention Systems (IPS) or Intrusion Detection Systems (IDS) can be configured to monitor key ingress/egress points and critical segments of the network to alert and/or block attacks as they occur:

- IPS – actively interfering with unwanted traffic
- IDS – merely monitoring the traffic and alerting or reporting on observed anomalies or known threats

There are different types of IPS/IDS as well as different models for deployment to be considered, when planning to implement monitoring of the environment. The main types of IPS/IDS include (Scarfone and Mell 2007):

- Network-Based: monitors network traffic across particular network segments or devices and analyzes the network and application protocol activity to identify suspicious activity. The IPS/IDS simply applies predetermined (mostly vendor provided) rule set to detect anomalous traffic.
- Network Behavior Analysis (NBA): examines network traffic to identify unusual traffic flows that may represent threats, such as distributed denial of service (DDoS attacks), certain forms of malware, and policy violations. The IPS/IDS first learns typical network behavior and then reports deviations from the baseline.

- Host-Based: monitors the characteristics of a single host and the events occurring within that host for suspicious activity.

Network-based IPS/IDS is generally the most prevalent type of monitoring, but host-based monitoring is a rapidly improving technology and provides detection of a broader range of internal and external attacks. The host-based IPS/IDS are typically bundled within a suite of tools in addition to e.g. antivirus. When planning to deploy an IPS/IDS system, the deployment model plays a key role in determining what the system can provide and how it affects the entire environment. In a traditional IDS role, the system is placed in an out-of-band mode (Fig. 5.4), whereby a monitor port or tap is copying all traffic to the monitoring system. IDS analyzes the traffic to detect an intrusion/attack. In this scenario, a possible attack generates an alert to a console or response team that responds to mitigate the attack and address the vulnerability that lead to the attack. This model is used in environments with concerns over blocking legitimate traffic, and the possible impact an inline device may have on flow of legitimate traffic.

In the Intrusion Prevention Mode all traffic is routed directly through the IPS system, inspected, then sent to the destination host/network (Fig. 5.5). In this scenario, traffic designated as an attack is dropped directly inline and is not sent to the destination. This model of implementation can potentially affect performance of legitimate traffic leading to careful evaluation of the placement prior to deployment.

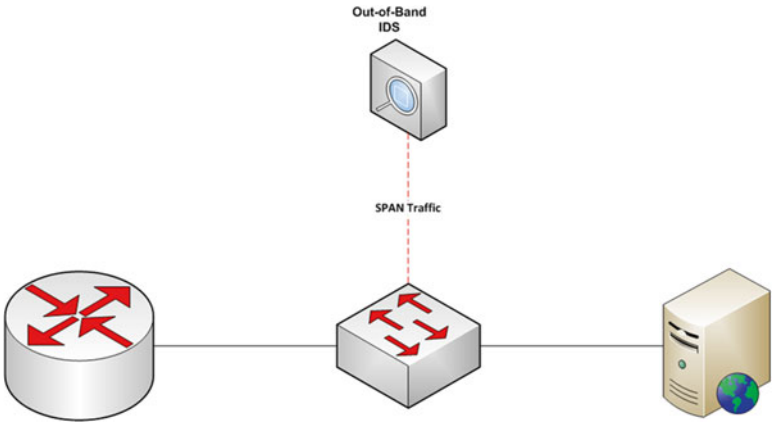


Fig. 5.4 Out-of-band intrusion detection system

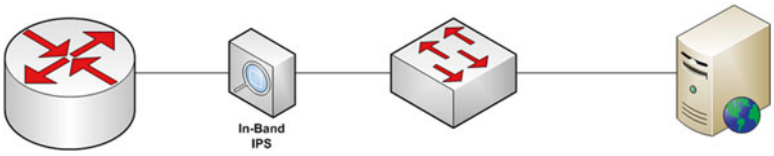


Fig. 5.5 In-band intrusion prevention system

However, in terms of attack mitigation, IPS provides the quickest response and most effective way of detecting and responding to attacks.

When planning to deploy an Intrusion Prevention System, options are available that can help address the typical concerns. The key options include:

- Performance: When purchasing/building an IPS, the network must have enough bandwidth to handle the amount of traffic that will be forwarded through the device. A number of vendors have technical specifications published for appliances, but purchasers must test the inline inspection using traffic simulators to ensure the system can handle the volume of traffic it will be inspecting in their institution. In any case, based on the IPS model, number of deployed rules, level of inspection and traffic patterns the impact on the traffic flow can be significant with possible business impact.
- Network Taps: When placing the system inline, bypass network taps must be purchased or built so that systems can be taken offline without affecting traffic flow in the environment. In maintenance windows, bypass taps can be put in bypass mode allowing reboots, updates, etc. without affecting the traffic in the environment.
- Placement: An IPS deployment for every network segment within an environment would be very costly. When planning to implement IPS, the institution should plan to place it in critical segments or ingress/egress points to get as much coverage as possible, without being required to deploy multiple devices to get complete coverage of the network. Systems placed within access infrastructure and key uplinks to routers/switches can give coverage across the environment, while not requiring multiple expensive devices to cover all segments of the network.
- Tuning/Configuration: Each vendor provides a set of signatures (configuration rules to identify bad behavior). These signatures are updated regularly to reflect the emerging threats. A number of implementations fail because signatures have been disabled, tuned out of use, or removed. Most vendors have so called 'Low-Fidelity' or 'High-Accuracy' signatures, which accurately distinguish between legitimate and illegitimate traffic, thereby minimizing false-positives that may block legitimate traffic. At a minimum, high accuracy signatures must be enabled for true blocking, whereas less accurate signatures are enabled for detection only. The number of enabled signatures might affect the performance of the IPS due to additional processing needed to match the signatures to the network traffic.
- Spend Time Tuning: When deploying the system, it is critically important to spend time developing a baseline to learn what is normal traffic in the environment, i.e. tuning the system. Ad-hoc configurations and deployments without adequate tuning can lead to a situation where signatures that have not been properly tested must be disabled and are not monitoring the environment. Adequate time spent tuning and building the signatures will lead to more effective monitoring and blocking.

Once an IPS is in place and properly tuned, it can provide a wealth of knowledge and visibility into the environment. In networks with high traffic flow, inevitably

both true and false alerts will be generated. Alert fatigue may occur. True alerts of attacks may be ignored or missed in the myriad of alerts that are generated by the IPS. To address this, a plan for monitoring and responding to alerts must be developed before the IPS is implemented. Unless there is a 24/7 presence of staff within the environment, organizations should evaluate the possibility of using a Managed Security Services Provider (MSSP) to monitor the system and alert staff when an attack is detected. An MSSP augments current staff and will help them gain knowledge of current trends. Such knowledge can be helpful in identifying actual attack traffic and reduce the amount of time spent monitoring and responding to attacks.

IPS/IDS systems complement firewalls. They allow monitoring of traffic that has been authorized for possible malicious content and facilitate responding to attacks in a timely manner. Properly deployed and configured within the environment, IPS/IDS provide additional layers of protection for the institution, which handles sensitive information and is potentially liable for breaches of security. Many modern firewalls have IPS/IDS built in and presented as part of the Unified Threat Management (UTM).

5.2.3 Remote Access/VPN

Translational research is inherently collaborative with both researchers and data frequently located at multiple geographically separate sites. Researchers need remote access to multiple networks. An infrastructure to support remote access must be implemented to allow users to access the private environment securely and reliably. There are many methods of remote access, ranging from traditional Virtual Private Networking (VPN) to other methods such as Remote Desktop/Terminal Services. Advantages and disadvantages of each must be considered when planning the infrastructure to permit remote access. It is important to take into account all the different methods that are going to be used/required/supported for remote access to the environment. Given the data-centric model for the application infrastructure, all access to applications can be considered ‘remote’ since there is no internal network. Looking at remote access in this way is valuable, when deciding what is going to be supported in the environment. To begin, planners must inventory:

1. the applications within the environment that are going to be accessed;
2. how users will access the applications; and
3. how administrators will access the applications.

Most recently developed applications utilize World Wide Web (www or web services), or have some form of web service client that accesses the application. Given this, some form of web remote access will be required and is generally provided through web reverse proxy servers or some other form of remote access or gateways. Administrators of applications may require access to terminal services or to some other type of client/server application. Table 5.1 can guide an inventory of the type of remote access that is going to be required for the environment.

Table 5.1 Methods of remote access by various applications

Application	Remote user access	Remote admin access
Public site	http/https	DB Server Studio, https
eCommerce site	https	https
eMail	https, eMail Client (MAPI)	VDI, https

Once the applications and their methods of remote access have been established, a proper set of remote access tools can be selected and implemented. When looking at remote access tools, remember that more than one method will generally be required. Not all applications, users and security models for remote access are going to work for all applications and users in the environment. It is important to select a few options that will work for most of the environment, then work to integrate the few that do not work into what is supported. Table 5.2 lists several options/methods for remote access along with their typical use, advantages and disadvantages.

Once methods for remote access have been selected for a particular environment, administrators can focus on the tools necessary to support all users and can configure them properly. Proper configurations include:

- Limited Access: Users should be restricted to those applications and services for which they have been pre-authorized. The default 'any' or 'allow all' methods of access should not allowed in the environment with an exception of administrator access well protected through a VPN. Different institutional policies are generally required for different classes of users.
- Logging/Monitoring: Logs of access by users regardless of method must be in place. Syslog and other Security Information and Event Management (SIEM) tools can be used to collect and analyze the remote access logs so that they can be generated when needed.
- Central Authentication: Ensure the remote access infrastructure uses a common identity format so that users are not required to remember multiple usernames/passwords and/or tokens to access the environment. Inconvenient access can encourage people to share credentials or methods of remote access, which can lead to elevated access privileges and violation of institutional policies.
- Multi-factor Authentication: Passwords are susceptible to multiple attack vectors including email phishing. Organizations are encouraged to add a second factor to the authentication process for remote access and highly sensitive servers. The second factor is typically in the form of a hardware token, mobile app token, text or call confirmation.
- Encryption: While encryption: While encryption of data in flight within a local network might be optional based on the network inspection needs and level of other mitigating controls – traffic through external networks is typically encrypted by default.

A well-conceived strategy for remote access with methods supported by the enterprise can provide access to internal systems and applications in a secure manner. Systems that are convenient for users permit them to perform their jobs in a secure environment, while ensuring that the institution complies with federal and state laws, regulations, and policies that govern access to sensitive data.

Table 5.2 Targeted use, benefits, and issues with various methods of remote access

Method	Target use	Benefits	Issues
VPN (IPSEC)	Full network access to the environment	Is mature and been around a long time. Is built into most security and firewall devices.	Can be problematic with slow/unreliable connections Gives user full network access (IP access) to the environment Network ports may not be opened from some secure networks
VPN (SSL)	Full network access to the environment	Uses standard https protocols Reliable across slow/unreliable network connections Open from most environments	Gives user full network access (IP access) to the environment
HTTP	Web-based applications over public and private networks	Browser-based Familiar to most users Has handlers/methods for building security onto protocol	Can be insecure in default deployments Very visible attack footprint
VDI (Virtual desktop infrastructure)	Access to a selected application or individual desktop	Thin client – small foot print, low bandwidth requirements unless video intensive applications. Security controls.	Complex setup, compatibility, performance for higher latency networks.
Terminal services/remote desktop	Full console access to remote servers and devices	Traditional desktop experience Full access to remote system Built-into Windows operating systems	Can be insecure in default deployments Older tools may not support authentication/encryption requirements Can give elevated access to systems Is not built for open/public networks using default configuration
Secure shell (SSH)	Remote user login to shell (linux) Remote user tunnel	Secure transport protocol for open/public networks Low cost	Takes more configuration from client side Requires configuration changes on some client machines

5.3 Operating Systems and Cybersecurity

Operating systems of applications are probably second to networks in terms of implications for security of the environment. Operating systems of servers are complex, and. have vulnerabilities that can lead to compromise of applications. It is important to choose operating systems that meet institutional needs, secure them by default, and continuously monitor and improve the security of the operating systems within an institution's particular environment. This section reviews challenges of securing operating systems with specific examples of how these can be addressed.

5.3.1 Configuration

Configuration standards, policies, and processes document the setup steps and procedures for configuring applications, network devices, servers, etc. within the environment. Configuration standards should be documented for each application or device within the environment. This ensures that applications and devices are consistently setup, securely built based on best practices, and are accessible in the event that primary support personnel are not available.

Initial configuration standards must be built for a particular environment. They should include:

- Best Practices: Do not re-invent the wheel. There are a number of sources of standards that have been well tested and can serve as a starting point for an institution (Scarfone and Mell 2007; Quinn et al. 2015; CIS 2016; NIST-NVD 2016).
- Do Not Copy Standards: Many of the available standards are purpose-built for specific regulatory and/or security requirements and may not be fully applicable to your particular organization and its environment. Institutions should use the standards of others as a guide, but standards adopted by an institution must fit its specific needs.
- Conduct Regularly Scheduled Reviews: Once a standard is written, it needs to change over time as technology, regulations, and the environment change. Configuration standards should be reviewed at least annually to make sure they are still relevant and accurate.
- Auditable: Compliance with standards must be auditable. This can be done by manual, scripted, or automated reviews. The Security Content Automation Protocol (SCAP) provides specifications and practical guidance to enable automated vulnerability management and audits of compliance (Quinn et al. 2010; NIST-SCAP 2016). Furthermore modern vulnerability scanners can validate setups (such as OS, database, applications) against published standards such as Center for Internet Security – CIS, Defense Information Systems Agency – Security Technical Implantation Guides (DISA STIG), etc.

There are a number of good sources for configuration standards and checklists. Examples include:

- Center for Internet Security (CIS)
  - http://benchmarks.cisecurity.org/en-us/?route=downloads.benchmarks
- NIST National Vulnerability Database Configuration Checklist Project
  - http://web.nvd.nist.gov/view/ncp/repository

In addition to these vendor neutral databases, most vendors provide guides to setting up the operating system with proper levels of security. Important items to consider, if a vendor guide is not available or is incomplete include:

- Services/Daemons: Disable all unnecessary services, programs, applications, and daemons that are not explicitly needed by the operating system. This reduces the footprint of the server and reduces the possibility of unneeded software affecting the operating system.
- Limited Access/Authorizations: Only allow local access to those administrators and power users who must directly access the system. This can be achieved through local authentication/authorization requirements along with network level controls for limiting access to administrative ports and services. This might also include disabling or renaming local accounts, changing default passwords, etc.
- Secure Management: When allowing remote management, insecure management protocols such as telnet, rsh, VNC, etc should be disabled and only secure management protocols enabled. This will reduce the possibility of individuals intercepting credentials of administrators and using them for malicious purposes.
- Auditing/Logging Events: System should be configured to log access and changes to the system. This will provide an audit trail of who made what changes and when. This can be useful in investigations of possible breaches of security and in troubleshooting.
- Patching/Updating: Each organization should build requirements into the standards to require periodic patching and updating of systems to ensure they have the latest software and any vulnerabilities that have been identified are patched as soon as possible.

When writing configuration standards, they should not be limited to operating systems and other network based devices, but should include any applications within the secure environment. This helps ensure consistent setups, knowledge transfers, and continued secure builds of applications running within the environment.

5.3.2 *Patching*

Given the size and complexity of applications and operating systems, it is inevitable that vulnerabilities will be found and require patching. This is highlighted in the latest Symantec Internet Threat Report (Symantec 2015), which noted a 30% increase in the number of vulnerabilities in software releases in 2010 compared to 2009. Patches are released to address vulnerabilities, after a period of delay while methods of correction are developed. If left unpatched, the vulnerabilities can lead to compromise of the system. To ensure these patches are applied and the systems are secure from these vulnerabilities, a patching program that includes regular scanning and updating needs to be put in place.

The first step in implementing a patching process is to regularly scan the environment for potential required patches. The scans should not be limited to server and applications, but should also include the network equipment (switches, routers, firewalls, etc) within the environment. Vendors generally release patches and updates for their equipment. Systems such as Windows and Linux have automatic update capabilities that can download and install the available patches for the system. There are products that automate scanning to identify vulnerabilities and available patches. Examples include Windows Server Update Services (WSUS) (Microsoft-WSUS 2016), Freeware OpenVAS (2016), Microsoft Baseline Security Analyzer (MBSA) (Microsoft-MBSA 2016), and Flexera Personal Security Inspector (PSI) (Flexera 2016). Once the list of required patches is developed and the available patches are downloaded, they must be installed on all the affected systems. As verification that patches have been installed properly within the environment, verification scans should be run against all the systems. This can be part of the follow-on monthly scans to ensure continuity and continued updating of systems within the environment.

The real challenge to implementing a patch process is setting a standard and regular process for installing patches. Most vendors have moved to a once a month release of patches (outside of critical updates). Organizations can follow the same schedule and regularly install patches on all systems, applications, and network devices within the environment. This will ensure the continued update and review of all the systems within the environment.

The best practices of patch management suggest a careful selection of patches to apply and apply them in succession to test environment before production for validation and avoid interference with production application. This is not always feasible given the staffing levels and amount of patches being issued by various vendors.

The inevitable argument that is raised is: ‘This will cause issues with my [system, application, device]’. If a patch or update has been shown to have compatibility issues with a particular application or device, there must be a process to document the exception, approve it, and regularly review it. The documentation should include the source of the exception (from support, through testing, etc), who generated the exception (owner), and the expected timeframe for resolving the problem. During

the regular scanning and installing of patches, all exceptions should be reviewed to ensure they are still applicable and required within the environment. Failure to resolve exceptions can lead to accrual of systems and applications within the environment that are out of standard and create potential vulnerabilities that could adversely affect the rest of the environment.

5.3.3 Vendor Management and Evaluation

In the fast paced world of information technology and cyber security many new vendors emerge every day with solutions that are often appealing from the execution, cost, feature and other perspectives. Organizations usually develop institutional standards for vendor and vendor products evaluation to assess conformance with industry standards, attention to security during the development and implementation, long term business viability, customer service, etc.

At minimum the organization should ensure that:

- vendor supported systems have a set schedule and that is agreed to upon contract/support agreement signing
- vendor is held accountable to appropriate standards

5.4 Protecting Sensitive Information

Sensitive information such as Personal Health Information is commonly used in translational research and must be protected. Typically, it is captured and stored in three types of systems: Research Patient Data Warehouse, tools for Electronic Data Capture that may include data storage, and Research Data Storage systems.

5.4.1 Protected Health Information

Protected Health Information, also known as Personal Health Information, (PHI) does not include publicly available information that is lawfully made available to the general public from federal, state, or local government records. Protected health information that cannot be exposed to individuals without permission of the patient is defined as information that when used alone or in combination with other information can identify an individual (whether living or deceased). Table 5.3 lists these identifiers, as defined in the United States by the Health Insurance Portability and Accountably Act (HIPAA) Privacy Rule.

Table 5.3 List of 18 HIPAA PHI identifiers

	Identifier
1.	Names
2.	All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code, if according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
3.	All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, and date of death and all ages over 89 and all elements of dates (including year) indicative of such age (except that such ages and elements may be aggregated into a single category of age 90 or older)
4.	Telephone numbers
5.	Fax numbers
6.	Electronic mail addresses
7.	Social security numbers
8.	Medical record numbers
9.	Health plan beneficiary numbers
10.	Account numbers
11.	Certificate/license numbers
12.	Vehicle identifiers and serial numbers, including license plate numbers
13.	Device identifiers and serial numbers
14.	Web Universal Resource Locators (URLs)
15.	Internet Protocol (IP) address numbers
16.	Biometric identifiers, including finger and voice prints
17.	Full face photographic images and any comparable images
18.	Any other unique identifying number, characteristic, or code (excluding a random identifier code for the subject that is not related to or derived from any existing identifier)

5.4.2 Research Patient Data Warehouse

The Research Patient Data Warehouse (RPDW) is one of the most highly protected components of the IT infrastructure in a research-intensive medical center. Typically it contains information from the electronic medical records of large numbers of patients. See Chap. 6, Research Patient Data Warehousing, for an extended discussion of data warehouses. Protection of PHI within a RPDW requires special attention to:

- Access control
- Auditing of control and data changes
- Strict change control of any alteration of the system

A RPDW typically contains huge amounts of data, e.g. patient's demographics, diagnostic and procedure codes, medical history, results of laboratory tests, images, and genomic data. Accruing, storing, and searching the data in a timely manner

requires high performance systems. Security systems must be carefully designed to avoid overly compromising performance of the warehouse. It requires a balance between ability to process and deliver data in a timely manner and ability to implement security controls that adequately protect the data. Performance can be improved by moving some controls to auxiliary systems surrounding the RPDW:

- Increased physical security of the data center as a trade off for on-disk data encryption
- A dedicated network segment surrounded by a firewall as a trade off for an in-band firewall
- Accessing data through an application server with a tight control rather than a direct database connection
- Application server to database connection might not be encrypted for performance reasons putting an additional burden on security of the application server
- The audit trail might be stored externally rather than stored on the system holding the RPDW

Many of the trade-offs are specific to an institution, its environment, and infrastructure. Institutional specificity might be related to:

- Processes for change control
- Types of queries or in general how the system is utilized
- The process of loading data (nightly full or incremental refresh, continuous/online versus batch load)
- The system of data backups such nightly, online DR, virtual vs. physical tape, etc.
- Method of user access (accessing a single RPDW or accessing a separate datamart that contains an extract of the primary data in a separate database)
- Reporting utilities and the security of service accounts accessing the data
- Utilizing all database built-in security controls – row level security, encryption

5.4.3 Tools for Electronic Data Capture

Tools for electronic data capture are commonly provided to investigators as part of the design and implementation of translational research projects. These allow researchers to collect data to be used in studies, clinical trials, etc. From the point of view of those responsible for security, it is important to note that tools for electronic data capture are typically user facing and hence require strong identity and access management. One such system might host a few or 100 s of studies making it a high level target for unauthorized users. In order to minimize the possibility of data compromise:

- Encryption between the client and the server is a must. This is typically done by SSL (https), but additional layers such as VPN or SSL VPN might be used

- Regularly scheduled changes in passwords must be enforced, typically at 90 day intervals
- Role management – distinguish between study coordinators and data entry personnel. Grant permissions based on need and role.
- Entitlement reviews – ensure that only the people who are entitled to access to particular data sets or studies receive access, and regularly review these entitlements.
- Comply with best practices such as regular server patches, centralized review and storage of the logs, change and configuration management.

5.4.4 Research Data Stores

Biomedical research institutions typically house many data generating groups/cores. In order to share data efficiently, network attached storage (NAS) or online (web) based storage systems are generally used. Much of the data in a medical research institution will contain Protected Health Information, as previously discussed for Research Patient Data Warehouses. There are a several basic considerations, when planning to protect information. In a NAS environment it is important to realize that;

- The data in flight might not be encrypted. The Common Internet File System (CIFS) does not have encryption built-in, whereas the Network File System (NFS) does in version 4 (IAPS 2007). Direct Internet access to file servers should be provided only through VPN.
- Although users can manage permissions to some degree themselves in CIFS and NFS environments, the built-in permissions management might not be ideal. Add on products and careful monitoring of permissions is encouraged in order to keep the permissions well-defined and manageable long term.
- Identity within the NAS is only as strong as the institution's identity management system. Inadvertent additions of users to groups, confusion of users with the same or similar names, unclear definition of groups and the data to which members have access are some of the issues that can plague an identity management system and must be clearly resolved. It is important to develop standardized workflows for addition and deletion of users and groups.
- Audit-trails are essential not only for compliance with laws and regulations, but as a best practice. Many users fail to realize the side effects of, for example, moving or removing data directories. Having the ability through audit trails to quickly identify who did what, when and where helps dispel many misconceptions and improve trust in the system.
- Data classification
- Data storage guidance (what goes where)

Web based data stores typically handle internet access better that NAS, yet can be mounted as a network drive e.g. using WebDAV (Web Distributed Authoring and

Versioning) protocol. With regard to security, some characteristics of web based data stores are:

- Encryption using SSL (https)
- Authorization/permissions are not limited to pre-defined protocols, but are dependent on the hosting application.
- Audit trails are easy to add
- During the upload process, information can be automatically parsed from the data and hence allow more flexible searches
- Typically stores "structured" data

5.5 Role of Users in Protecting Information

The focus of this chapter on institutional cybersecurity has been on technology. In reality, breaches of security with loss of protected or confidential information usually occur because users fail to comply with institutional policies and standards rather than because of technical failures of network security and intrusion prevention systems. For example, theft, loss, and/or misuse of portable devices are one of the most common breaches of security across industries (Verizon 2011). It is critically important that institutions develop and enforce policies defining acceptable and unacceptable uses of information resources. Institutional policies and standards should address such issues as:

- Behaviors of users must be consistent with the mission, vision, and core values of the institution.
- Users must comply with applicable laws, regulations, and policies.
- Users acknowledge that sharing of Confidential Information with unauthorized individuals may result in the waiver of legal protections and may place the user and employer at risk of criminal or civil liability or may damage their financial standing, employability, privacy, or reputation.
- Users are individually responsible for maintaining the confidentiality, security and integrity of their chosen passwords.
- Users must promptly and properly dispose of, sanitize, and/or destroy confidential information in any form (e.g., paper, electronic, on Portable Devices) when no longer useful
- Users consent to institutional auditing and monitoring for excessive or inappropriate personal use, criminal activity, regulatory violations.
- Users understand that violations of any policy may subject them to disciplinary action up to and including termination of employment.
- Users must ensure that: Confidential Information is not visible in an unattended work area or on an unattended workstation; they log off of systems after access; they do not post passwords or user IDs in visible areas.
- Users are responsible for ensuring optimum security of Portable Devices (regardless of device ownership) that access, transmit, or store the institution's

Confidential Information. Users must immediately report any loss or theft of computing devices.
- Users are prohibited from interfering with the intended functioning of the information system, e.g. disabling security devices, inserting viruses, inserting network monitoring devices, installing unapproved software, destroying files.
- Security awareness training – having system users understand the types of threats that they could fall victim too, such as social engineering and phishing campaigns and the risks associated with them.

To summarize, maintenance of high levels of cybersecurity to protect confidential and protected health information requires a combination of best practices in technology and thoughtful institutional policies that are enforced. One without the other precludes success.

A number of frameworks have been developed to aid organizations with risk assessment and mitigation. Our organization is subject to FISMA (Federal Information Security Management Act) for a number of federal contracts. FISMA uses NIST 800-53v4 (JOINT-TASK-FORCE 2013) "Security and Privacy Controls for Federal Information Systems and Organizations" and we follow NIST standards for the annual Information System Security Plan (ISSP).

5.6 Joint Management – Hospital and Research

As discussed in Sect. 4.8, to support research in a research oriented medical center, it is important to define clearly who is responsible for providing IT support to the different missions – patient care, business operations, and research. In the authors' institution, clinical and business operations are supported by hospital information services (IS), while research is supported by biomedical informatics (BMI). The two groups are separately staffed and budgeted, but must collaborate very closely to provide coherent support to the overall research and clinical enterprise. With regard to operation of networks and implementation of cybersecurity, IS and BMI have developed a responsibility matrix (Table 5.4). This has served the organization well and prevented misunderstandings. IS is responsible for installing and maintaining the physical networks, including firewalls. BMI works with IS to design networks

Table 5.4 Distribution of responsibilities for support of research between Hospital Information Services (IS) and Biomedical Informatics (BMI)

Issue	IS	BMI
Physical networks	X/J	J
Security	X/J	J
Authentication	X/J	J

X primary responsibility for operational support, *P* primary responsibility to formulate policies, *J* participates in formulation of policy

that support research programs and to develop standard operating procedures for identity and access management. The goal is to create an environment in research that is secure and meets regulatory requirements with regard to protection of sensitive information. With this in place, the hospital can transfer clinical information from electronic medical records to the research data center with assurance that the information will be properly protected.

References

CIS. Center for internet security. 2016. Retrieved March 21, 2016, from http://www.cisecurity.org.

Cruse A. Processor privilege levels. 2016. Retrieved March 21, 2016, from http://cs.usfca.edu/~cruse/cs630f06/lesson07.ppt.

Flexera. Flexera personal security inspector. 2016. Retrieved March 21, 2016, from http://www.flexerasoftware.com/enterprise/products/software-vulnerability-management/personal-software-inspector/.

IAPS. NFSv4: overview of new features. 2007. Retrieved March 21, 2016, from http://www.iaps.com/NFSv4-new-features.html.

JOINT-TASK-FORCE. Security and privacy controls for federal information systems and organizations, NIST special publication 800–53 Revision 4. Gaithersburg: U.S. Department of Commerce, National Institute of Standards and Technology; 2013.

Microsoft-MBSA. Microsoft baseline security analyzer. 2016. Retrieved March 21, 2016, from http://technet.microsoft.com/en-us/security/cc184923.

Microsoft-WSUS. Windows server update services. 2016. Retrieved March 21, 2016, from http://technet.microsoft.com/en-us/updatemanagement/default.aspx.

NIST-NVD. National vulnerability database checklist program. 2016. Retrieved March 21, 2016, from http://web.nvd.nist.gov/view/ncp/repository?tier=&product=&category=&authority=&keyword=.

NIST-SCAP. Security content automation protocol (SCAP). 2016. Retrieved March 21, 2016, from https://scap.nist.gov/.

OpenVAS. OpenVAS: open source vulnerability scanner and manager. 2016. Retrieved March 21, 2016, from http://www.openvas.org/.

Quinn S, Scarfone K, Barrett M, Johnson C. Guide to adopting and using the security content automation protocol (SCAP) version 1.0, NIST special publication 800–117. Gaithersburg: U.S. Department of Commerce, National Institute of Standards and Technology; 2010.

Quinn S, Souppaya M, Cook M, Scarfone K. National checklist program for IT products – guidelines for checklist users and developers, NIST special publication 800–70 Revision 3. Gaithersburg: U.S. Department of Commerce, National Institute of Standards and Technology; 2015.

Scarfone K, Mell P. Guide to intrusion detection and prevention systems (IDPS), NIST special publication 800–94. Gaithersburg: U.S. Department of Commerce, Technology Administration, National Institute of Standards and Technology; 2007.

Squid. Squid cache proxy project. 2016. Retrieved March 21, 2016, from http://www.squid-cache.org/.

Symantec. Symantec threat report. 2015. Retrieved March 21, 2016, from http://www.symantec.com/threatreport/print.jsp?id=threat_activity,vulnerabilities,malicious_code,fraud_activity.

Verizon. Verizon data breach report. 2011. Retrieved March 21, 2016, from http://www.verizonbusiness.com/resources/reports/rp_data-breach-investigations-report-2011_en_xg.pdf.

Chapter 6
Data Governance and Strategies for Data Integration

Keith Marsolo and Eric S. Kirkendall

Abstract Recent years have seen a dramatic increase in the overall adoption of electronic health records and other ancillary systems in health systems across the United States. This has led to a corresponding increase in the volume and breadth of data that are generated during the course of health care operations. Institutions have a strong desire to mine this information for analytic, improvement and research purposes. Most efforts involve the integration of data from multiple source systems, transformation and synthesis into a consolidated view. Several common strategies exist, including the use of data warehouses or integrated data repositories, as well as the creation of project-specific data marts. We describe several of the most common data models used to when creating integrated data repositories for research purposes, as well as the basic steps required for implementation. We also discuss the importance of data governance in healthcare, which includes the tools, policies and procedures that ensure data are used effectively within an institution. This includes efforts around data characterization and data quality, the use of data stewards, management of metadata, and more.

Keywords Common data models • Data governance • Data integration • Data warehousing

K. Marsolo, Ph.D. (✉)
Departments of Pediatrics and Biomedical Informatics, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Avenue, MLC-7024, Cincinnati, OH 45229, USA
e-mail: keith.marsolo@cchmc.org

E.S. Kirkendall, M.D.
Departments of Pediatrics and Biomedical Informatics, Divisions of Hospital Medicine and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine,
3333 Burnet Avenue, MLC-3024, Cincinnati, OH 45229, USA
e-mail: eric.kirkendall@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_6

6.1 Introduction

The increased adoption of electronic health records (EHRs) in the United States over the past few years has led to a tremendous increase in the amount of data that are captured during the course of clinical care and has stimulated efforts to ensure that information is put to "meaningful use" (Blumenthal 2009, 2010, Blumenthal and Tavenner 2010). This increase has led to a corresponding desire to analyze and mine that information for research and quality improvement purposes (Weiner and Embi 2009). Despite the increased implementations of enterprise EHRs, data often exist in multiple disparate systems. As such, there is a need for systems that support the integration of data from multiple sources and of multiple types into a coherent view, which can then be used for decision support, analysis, and reporting purposes. Initial efforts within healthcare have focused largely on the creation of data warehouses (also called integrated data repositories) (Mackenzie et al. 2012), which have been employed by the business community for decades to integrate and analyze financial data (Kimball 1996, 1998). Because of cost to develop and maintain an enterprise-wide data warehouse, and the complexity of integrating specialized data sources, recent years have also seen the spread of more single-purpose solutions, such as project- or content-specific datamarts. These marts may be focused on a single condition, such as heart disease, and may be designed with a single purpose in mind, such as reducing cost or eliminating variations in care. Instead of there being a single solution to the problem of data integration, they exist on a spectrum, with specific architectures optimized to meet specific needs. The decision on which option to choose will depend on the project. The fact that there may not be a single, physical source of truth for all things data within an organization also indicates why data governance is an increasing area of focus for many healthcare institutions.

6.2 Spectrum of Clinical Information Systems

Clinical information systems exist on a continuum within the healthcare delivery environment (Fig. 6.1). EHRs, by their nature, are patient-centric, allowing users to interact with the record of a single patient, one patient at a time. They are designed to be high-availability, and as such, are typically transactional in nature. This results in an application and data model that makes it very difficult to perform population or cohort-level analysis.

To offset these limitations, many EHRs are packaged with an operational data store (ODS), a database that can be used for reporting purposes. Most ODSs are still very transactional in nature, often reflecting the underlying architecture of the EHR (to simplify the process of refreshing the ODS with updated data). As the name implies, one of the primary uses of an ODS is to generate data that provides insight into the operations of an institution – clinical, financial, or otherwise. However, they are complex. For example, the ODS for the Epic EHR (Clarity) has over 15,000

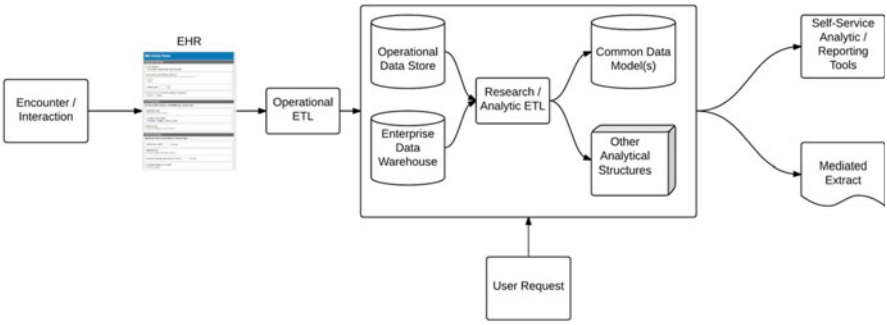


Fig. 6.1 Example data flow within a healthcare institution. Data are captured in the EHR during an encounter, and are then transferred to an operational data store and/or enterprise data warehouse using an operational extract-transform-load (ETL) process. A secondary ETL may then be used to populate one (or more) common data models that are to be used for research, as well as any other analytical data structures needed by the institution. Data are provided back to users through a mediated process or self-service query tool

tables. It is often difficult, if not impossible, to integrate information from outside sources into the ODS. In addition, population level analysis is still a challenge, as the relevant data may be spread out over dozens of different tables.

Integrated data repositories are designed to overcome the limitations of ODSs. Data are pulled from multiple tables/sources and combined into a single integrated record that can then be queried and analyzed. These repositories may exist as large, enterprise-wide data warehouses that incorporate clinical, financial and administrative data; they may be research-focused repositories that adhere to a common data model that allows for queries to be replicated or distributed across multiple institutions; or they may be tailored analytical structures that are designed to focus on a specific domain or condition. Research-specific or analytical-specific structures may be populated with their own specific loading procedures. Any one of these models may then be used to service user queries, either in a self-service fashion or through a mediated process. For the purposes of this discussion, we will focus on those repositories that are intended to support the research enterprise.

6.3 Research-Specific Integrated Data Repositories

To facilitate research and quality improvement, many institutions have or are now creating research-specific integrated data repositories (Eggebraaten et al. 2007, Lowe et al. 2009, Murphy et al. 2010, Zhang et al. 2010, Mackenzie et al. 2012). While it is possible to use an operational data warehouse for research purposes, a research specific-repository is intended to support a different set of use cases and workflows and will often include additional data sources (e.g., genomic, sensor data). In order to address privacy and security concerns, particularly those related to

the Health Insurance Portability and Accountably Act (HIPAA) Privacy Rule and Health Information Technology for Economic and Clinical Health (HITECH) Act (Title 45 Code of Federal Regulations, Blumenthal 2010), they typically add additional safeguards like de-identification or obfuscated access.

6.3.1 Use Case

For any research data repository, the initial step of almost every query is almost always some form of cohort identification, i.e., find patients who meet criteria X, Y and Z. Query criteria can include demographics (age, race, gender), diagnoses, laboratory results, procedure codes, medication orders, etc. and can be constrained by date, exclusion/absence or number of occurrences. While cohort identification may seem like a rather trivial use case, it is often the first step for any kind of biomedical analysis, as it is necessary to first define the population in question. This holds whether the query is to support clinical care, research or quality improvement (e.g., find all patients of a particular doctor or clinic, all patients with a given disease/genetic anomaly, all patients who have not come to clinic during the past 6 months). If the data are housed in a single warehouse, cohort queries can be complex, but are feasible. If the data are stored in a set of siloed source systems, they can become incredibly time-consuming, if not impossible, to complete.

Once an initial cohort has been generated, a data warehouse can be used to facilitate the analysis of data about its members. This can occur through tools that have been developed to sit on top of the warehouse and interact through a middleware service, as in STRIDE, i2b2, and Physio-MIMI (Murphy et al. 2006, Mendis et al. 2007, Lowe et al. 2009, Zhang et al. 2010), which allows for self-service or mediated access by super users (example below). Alternatively, the warehouse can be a source of cohort- or project-specific datasets, serving as the source of extracts or data marts that can be analyzed with a user's favored statistical package.

While cohort identification is typically the primary use case for an integrated data repository, other use cases include: searching for and requesting biosamples available for research (another version of cohort identification); augmenting it with patient outcome and process measures (which would be stored as derived observations), and as a data source for research and quality improvement registries.

6.3.2 Data Model

There are a number of different data models that can be chosen to underlie an integrated data repository. The simplest and most traditional is the star schema, which is employed by the Informatics for Integrating Biology and the Bedside (i2b2) framework (Murphy et al. 2006, 2010, Kohane et al. 2012). In this model, a central fact (observation) table is linked to a number of different dimension tables, which

are used to group related data (for example, by patient, visit, or provider). In a research data warehouse, observations can be anything from the presence of a diagnosis, a patient's birthdate, a lab result or a billing record. A slightly more complex version of this model is the snowflake schema, where a single fact table is linked to a number of dimensions that are normalized into multiple tables (Kimball and Ross 2002). Other alternative data models include the Health Level 7 – Reference Information Model (HL7-RIM) (Eggebraaten et al. 2007), the Observational Medical Outcomes Partnership (OMOP) Common Data Model (Reisinger et al. 2010), which is managed through the Observational Health Data Science and Informatics (OHDSI) program (Hripcsak et al. 2015), the Virtual Data Warehouse Common Data Model (Ross et al. 2014), Mini-Sentinel (Curtis et al. 2012) and the PCORnet Common Data Model (Califf 2014).

The choice of data model should be influenced by the by the primary use case of the repository. There is no single "best" model that is optimized for all research needs. All of them will require tradeoffs in one form or another. Some models, like i2b2 are designed to efficiently represent data and quickly identify cohorts, but require additional transformations to efficiently support downstream analytics. Others, like Mini-Sentinel and the PCORnet CDM (which is derived from the Mini-Sentinel model) are designed for analytic utility and to be easily understood the end user, but use inefficient table structures. The OMOP model, on the other hand, requires a larger degree of standardization when loading data, though that effort is rewarded with more simplified analytical queries. Groups have begun to create transformations that allow users to convert from one model to another, though due to differences in the table structure or terminologies, information may be lost. In addition, converting from the same source system to different data models may yield different analytical results (Rijnbeek 2014, Xu et al. 2015).

6.3.3 Terminologies

A particular challenge in creating integrated data repositories revolves around terminologies, or how to code a particular data element. There are two major approaches: either rely on a "local" code (typically derived from the coding scheme of the source system, though the source system may utilize a standard terminology for specific domains), or transform the data so that it is consistent with one of the many terminology standards, like ICD-9/-10 or SNOMED for diagnoses, LOINC for laboratory results, or RxNorm for medications. For data domains that may be represented by many different terminologies, such as diagnoses, which are often coded as ICD-9, ICD-10 and SNOMED, users may decided to standardize on a single terminology and represent all data of that type using that coding scheme. This is the approach behind the OMOP model. Terminologies are also discussed in Chap. 3 (Standards for Interoperability).

The decision to transform the data again depends on the primary use of the repository. If data will be routinely shared with outside institutions, transforming an element to a standard can ease the process of ensuring semantic interoperability. If the main users of the repository are primarily internal customers, it may be beneficial to retain a link to the coding used in the source system(s). Local users will be familiar with local terms that appear on the application screens of the source and can navigate more easily in the research system. While this decision often had to be a binary choice, many CDMs now include the option of storing the original source value as part of the observation, which allows the end user to verify the mapping and investigate the original coding, if desired.

One benefit to utilizing local source terminologies is that it may help eliminate confusion for end users. If an EHR uses ICD-9 as an internal coding for diagnoses, for instance, and the repository converts them to SNOMED, a query for all "asthma" patients run on the clinical ODS may not match the results returned from the research repository because SNOMED codes at a different level of granularity than ICD-9. ICD-9 is a relatively coarse terminology system, and converting to a fine-grained terminology like SNOMED can lead to errors if not done correctly. Discrepancies of this kind can lead to frustration on part of staff, as they seek reasons for the discrepancy in results obtained from searches of the two local repositories that are supposed to be populated from the same data sources. Such discrepancies can cause users to lose faith and trust in future results.

A particular advantage to using internal terminologies is that they are easier to maintain. If the source system changes the way that items are defined, the terminology of the warehouse can simply be updated to reflect the modifications. If items are mapped to a reference terminology, the work is essentially doubled. The internal terminology will need to be monitored for modifications, as will the reference terminology. While there are tools to help with mapping (Rubin et al. 2006, Musen et al. 2012), the process almost always requires time-consuming manual reconciliation. Government regulations like Meaningful Use are also spurring the adoption of standards within the EHR and other source systems (Blumenthal and Tavenner 2010), which should help eliminate the proliferation of local coding schemes, though it will take some time to propagate fully through the industry.

6.3.4 Data Provenance

Clearly understanding the original source of a data element (data provenance) within an integrated data repository is essential, but is a significant challenge. Within an EHR, there may be multiple diagnoses and each may come from a different source within the system (admission, discharge, problem list, encounter, billing, etc.). Depending on how the diagnosis was generated, some sources of the data element, may have more "credibility" than others. For instance, a diagnosis of lupus erythematosis assigned in an optometry clinic may not be as reliable as one assigned by a rheumatologist. Within an EHR or ODS, items like diagnoses are typically

stored in a location that is based on how the data is entered into the system. In an integrated repository, all of this information is brought together into a single table of observations. If care is not taken to denote where a value came from, or to allow users to add filters so that they can query on multiple parameters such as clinic, provider, and billing office, then results returned to users may not mean what they think it means. This becomes a concern particularly when users are allowed to generate self-service queries.

6.3.5 Extract-Transform-Load (ETL)

One of the biggest challenges in creating an integrated data repository lies in the procedures used to extract the data from the source systems, transform it to the correct format and load it into the target tables (Kimball and Caserta 2004). This process (extract-transform-load) is often referred to as ETL. Maintaining an ETL process can be expensive in terms of both people and software. Enterprise EHRs are generally supplied by commercial vendors and are constantly being updated with new versions. Organizations may also modify the local EHR to meet needs of internal users. Automated ETL tools like Microsoft SQL Server Integration Services (SSIS), Oracle Data Integrator, Informatica, and Talend can help monitor the quality of the source data and notify administrators of any changes. In many cases, though, manual SQL procedures must be created to optimize and customize the process.

The design of an ETL process depends on the frequency of the data refresh from the source systems, as well as the frequency of refresh to the target. Dealing with a laboratory information system (LIS) that sends results as real-time HL7 messages is a more complex endeavor than one that transmits results as a daily flat file. The closer to real-time the transmission of data from a source is, the more difficult it is to monitor and maintain the ETL process. The same rule applies when refreshing a target system. A challenge when dealing with EHR data is that the size of an ODS can be in the range of terabytes. A complete refresh of an integrated data repository with data of this size will eventually take longer than the available refresh window. This can be mitigated by the creation of an incremental or delta refresh, where the only data that are loaded are the new/changed elements. A refresh like this is more challenging to architect, as one must determine which elements have been added/deleted/modified, which can be difficult depending on the size and complexity of the source. While ETL processes to load EHR data into a repository can occasionally be shared among institutions, most EHR implementations include a fair degree of customization, requiring similar customization in the corresponding ETL process. If data are routinely pushed from the repository to other sources (data marts, registries, etc.), this adds complexity to the refresh process because a system must be in place to move data from the source to the repository to the users' application in a timely manner.

6.3.6 *Privacy*

Laws governing human research vary from country to country. This discussion focuses on requirements in the United States. Issues of privacy and consent are discussed in more detail in Chaps. 2 (Protecting Privacy in the Child Health EHR) and 5 (Institutional Cybersecurity in a Clinical Research Setting). In order to conduct research with patient identifiable patient data, users must obtain consent (Title 45 Code of Federal Regulations). The alternative is to work with a de-identified dataset, or a dataset that has been stripped of all personal health information (PHI). Working with de-identified data is not considered to be human subjects research and does not require an IRB protocol. This makes the creation of an integrated data repository for research purposes a challenge, as the act of creating the repository requires the use of identifiable patient data. Users of repositories that have been created for clinical or operational purposes do not face these challenges because the use of PHI in those contexts is covered under the HIPAA Treatment, Payment and Operations (TPO) exclusion. There are several strategies that can be employed to satisfy the requirement that consent be obtained from a patient in order to have their data included in a research warehouse. The first is simply to have each patient sign a consent form denoting whether their data can be included. This approach is relatively expensive and time-consuming, however, because of patient volume, and is almost never used. A variation of this approach is to have hospital registrars ask for patient consent as part of the registration process (Marsolo et al. 2012). Such a "self-consenting" process, i.e., a member of the research staff is not immediately available, may grow in favor as enterprise EHRs with integrated registration become more prevalent, and as the Department of Health and Human Services considers changes to the Levenson (2015). Another strategy, which has fallen out of favor in recent years, is to employ a passive consent, or "opt-out" strategy, where all patients and their data are included by default unless they explicitly request to have their information excluded. At this point, the most common approach is to request that an IRB waive the consent requirement, the justification being that it is impractical to obtain consent from every single patient at a hospital. The final method is to include language in the clinical consent-to-treat document that all patients must sign stating that any of their data may be used for research and included in a research patient data warehouse.

Because many warehouses are established under an IRB protocol that waives patient consent, they typically operate under the "honest broker" principle (Dhir et al. 2008). An honest broker serves as an intermediary between the individual providing the data and a researcher (SACHRP 2010). The broker aggregates and integrates identified data from multiple sources, removes the identifiers and then provides the resulting dataset to the investigator. They provide a service, but do not function as a member of the research team (i.e., they cannot participate in the analysis of the data). The same principle can be applied to biorepositories and biospecimens, as covered in Chap. 7, Laboratory Medicine and Biorepositories.

Table 6.1 lists the 18 identifiers specified by HIPAA as PHI and whether they can be included in a de-identified or a limited dataset. The traditional method of de-

Table 6.1 HIPAA identifiers and whether they can be included in a de-identified or limited dataset

Identifier	Included in de-identified dataset	Included in limited dataset
1. Names	No	No
2. Postal address information	The initial three digits of a zip code, if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000;	Yes
3. Social security numbers	No	No
4. Account numbers	No	No
5. Telephone & fax numbers	No	No
6. Elements of dates for dates directly related to an individual, including birth date, admission date, discharge date, date of death	No – except for year	Yes
7. Medical record numbers	No	No
8. Certificate/license numbers	No	No
9. Electronic mail addresses	No	No
10. Ages over 89 and all elements of dates indicative of such age	No	No
11. Health plan beneficiary numbers	No	No
12. Vehicle identifiers & serial numbers, including license plate numbers	No	No
13. Device identifiers & serial numbers	No	No
14. Web Universal Resource Locators (URLs)	No	No
15. Internet Protocol (IP) address numbers	No	No
16. Biometric identifiers, including fingers and voice prints	No	No
17. Full face photographic images & any comparable images	No	No
18. Any other unique identifying number, characteristic or code that is derived from or related to information about the individual	No	No

identification is through the HIPAA "safe harbor" standard, which states that if all of the 18 identifiers listed under the HIPAA Privacy Rule are removed and if there is an expectation that it is not possible to use the remaining data to identify a person, then the data are de-identified (the other method, Expert Determination, involves statistical de-identification and is not routinely used because of its complexity and requirements for certification). Investigators may use a de-identified data set in their research without IRB approval. A limited dataset is one where 16 of the 18 HIPAA identifiers are excluded, the exceptions being full dates and zip code. The use of a limited dataset requires that the investigator submit a protocol to the IRB and receive approval before conducting research using the limited data set.

In order to provide investigators with information on patient age more specific than a year (i.e. age in months, days or weeks), and to provide data on date of admission, discharge, or service, without requiring an IRB protocol for every dataset, one approach is to use "scrambled" dates. Before loading the data into the warehouse, one applies a date shift to each patient's record. For example, for a given a patient, all dates in the record are shifted backwards a random number of days between 1 and 180. The shift is consistent within a patient's record so that sequences of events are preserved, but each patient in a cohort is shifted by a different randomly chosen number. This makes the data unusable for surveillance research, but does allow for complete information about sequences of events to be released to investigators as part of a de-identified dataset. Depending on the institution, the use of this approach may require that the repository undergo the Expert Determination method of de-identification, demonstrating that the shift leaves the data de-identified.

6.4 The i2b2 Research Patient Data Warehouse

This section describes an example of a widely used research data warehousing platform and how it can be deployed within an institution. This example focuses on i2b2, an informatics framework developed by the NIH-funded i2b2 National Center for Biomedical Computing. One of the initial, motivating use cases of i2b2 was to serve as a bridge between clinical research data and the vast data banks arising from basic science research in order to better understand the genetic bases of complex diseases. i2b2 has been released as a collection of open source software tools that facilitate the collection and aggregation of patient data, both clinical and research, into a translational research repository (Kohane et al. 2012).

The basic function of the software is to allow the user to identify cohorts of patients from a large repository of observations. The core code manages projects, files, a query ontology, data and workflows. It is designed to allow developers to add functional modules, called cells, to the core code. Cells have been developed to extend the core to a multitude of activities like natural language processing,

annotation of genomic and image data, data analysis, concept mapping, and linkage to other research systems (Mendis et al. 2007, 2008).

The functional design of i2b2 is based around a central ontology of query terms consisting of standard and local terminology to describe clinical and research observations. The i2b2 ontology is fully extensible to local terminology sets and contains terms for diagnoses, procedures, demographics, laboratory and microbiology tests, visits, service providers, genomics, proteomics and imagery. Users select terms from the ontology in logical combinations (like a Venn diagram), and i2b2 generates a query against its underlying database to deliver cohorts of patients meeting the selected criteria.

i2b2 employs a data model based on the data-warehousing-standard star schema and runs on common database engines like Oracle and SQL Server. All of the patient observations are contained in a central database table. Each observation is linked to a concept which links back to the central ontology. More detail is provided in the following section.

Communication in i2b2 occurs through a middleware of web services. These services pass information between the different components, from the database layer up to the front-end application. Currently, users access i2b2 through one of two different approaches. One is by using a derivative of the open source Eclipse workbench that runs as a workstation-based client and the other is through a browser-based web client.

6.4.1 i2b2 – Data Model

The data model embedded in i2b2 uses a variation of the entity-attribute-value model, where a number of attributes, each with a corresponding value, are stored about an entity (patient). Attributes are facts, measures or observations about a patient, such as demographics, diagnoses, or laboratory results. It is also possible to define de novo attributes or create variables that are derivations or combinations of existing ones. Observations are tagged with one or more concepts, which can come from standard terminologies like SNOMED or LOINC or from custom definitions specific to a project. These codes are used when querying the warehouse. Collectively these concept codes are the items that can be used in a query. For example, when a user submits a query asking for all patients with a diagnosis of ‘asthma’, the i2b2 query engine will look for anyone with an observation that has the code associated with ‘Asthma.’ It is also possible to assign synonym codes, so that a search for ‘Asthma’ would return all patients with a corresponding ICD or SNOMED diagnosis.

6.4.2 i2b2 – Privacy Mechanisms

The i2b2 framework employs a number of measures to ensure the protection of patient identity and privacy (Murphy et al. 2011). Access in i2b2 is granted on a project level, with the ability to assign a different role and access level to the same user for each project. Users can be *administrators*, which gives them full control over a project, a *manager*, which allows them to see all the previous queries of the project's users, or simply a *user*, where all they can see is their own queries. In addition, i2b2 uses a number of different access levels to restrict the amount of data that a particular user can see. The lowest level of access is at the *obfuscation* level, where the user can only see the approximate results of an aggregate query. A Gaussian (of 1.7) is applied to the aggregate count and the values are returned with an accuracy of +/− 3. A user's account is locked if they try to run the same query more than a certain number of times, as doing so would allow them to try and zero in on the true aggregate value. A user with the *aggregate* level can see the exact aggregate counts, but not individual patient data. The *limited* role allows the user to see individual patient data, but only the information that is available in a limited data set (i.e. de-identified plus dates of birth and dates of service). Users who have access to a project at the *limited + de-identified* level can see all of the data in the limited dataset plus (presumably) de-identified text results, assuming they are available. Finally, users can be granted *full* access, with the ability to see fully identified patient records. Custom access levels can also be created if needed.

6.4.3 i2b2 Workbench

The i2b2 workbench allows users to execute self-service cohort queries, as illustrated in Fig. 6.2 (these screenshots reflect the workbench developed at CCHMC). The workbench includes a search tool (1) that can be used to find search terms within the ontology (2). These terms can be dragged and dropped into a into a Venn diagram-like interface (3). Upon execution (4), the query will return an obfuscated count of the number of patients that meet the specified criteria (5). If a user wishes to access their previous results, they can do so through the previous query section (6).

6.5 Data Governance

The descriptions of integrated data repositories in this chapter have focused on those intended for research use. At most, if not all healthcare institutions that are involved in research, there are operationally-focused efforts in areas such as quality improvement, analytics, clinical or financial decision support, predictive analytics, and so forth. The steps to prepare a dataset for these activities are essentially the same as those needed for research. However, for a variety of reasons, including

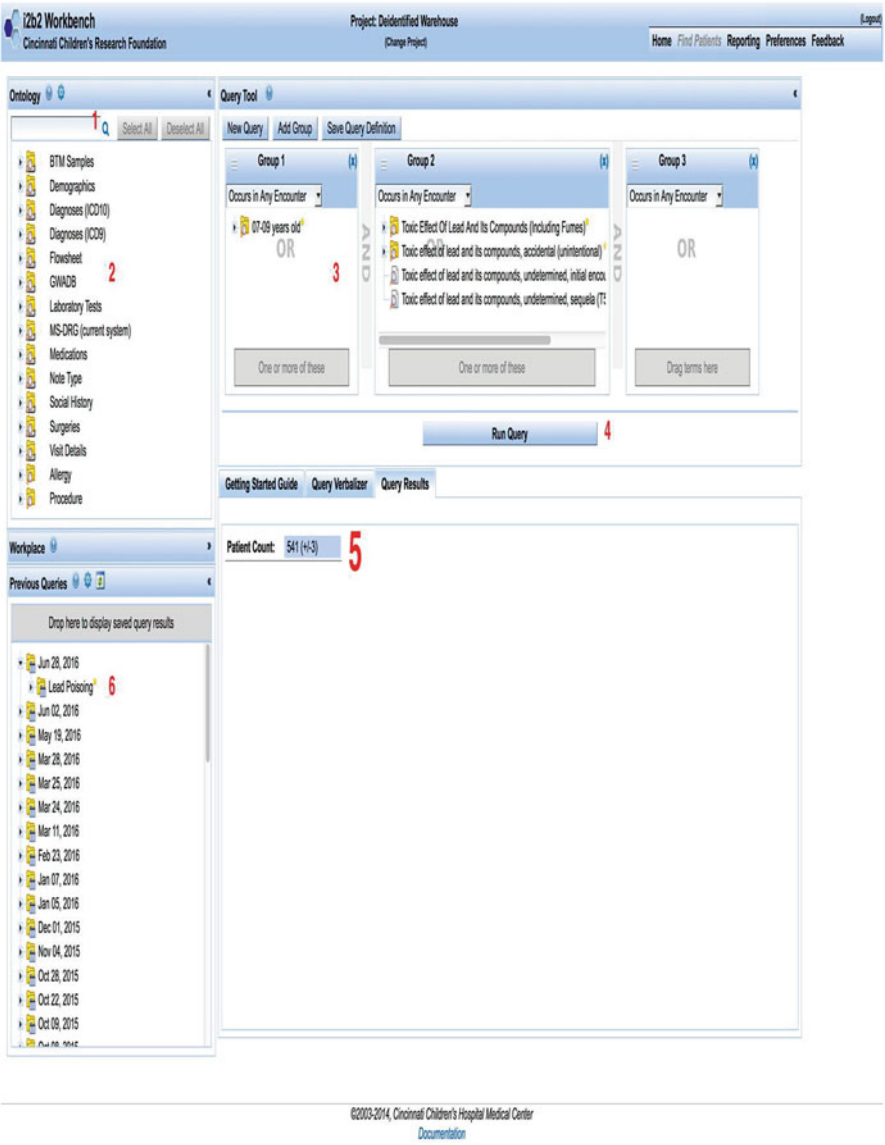


Fig. 6.2 Screenshot of an i2b2 query tool. Users can select inclusion/exclusion criteria in order to identify patient cohorts

technology, funding, personnel, operational, or legal reasons, they often happen in isolation from the rest of the institution. Even within the operational teams, the group making decisions related to quality improvement may not be in frequent communication with those involved in predictive analytics. As a result, there is growing recognition of the need for data governance within healthcare, in order to make sure that there is alignment in how decisions are made about data across an institution.

Data governance is one of the most important factors in how well an organization maximizes and uses its data assets. While many entities generate an abundance of data, not many take the necessary steps to effectively and efficiently manage the storage, cataloguing, assessment, cleaning, transformation, and provisioning of the data. Each of these activities (and others not listed above) requires additional effort that may constitute an upfront cost. However, the long-term benefits of performing these tasks will outweigh the initial investment if planned intelligently and in a manner that makes sense for the organization.

The exponential proliferation of data and data sources at healthcare institutions is exposing the suboptimal nature of traditional data management practices and highlighting the need for a more formalized approach. Emerging technological trends such as the Internet of Things, co-creation of data, RFID tagging, and the concept of big data in general are contributing to the health sector's recognition of the problem. If institutions do not take corrective actions, the growth of data will only serve to become an even bigger version of the '*garbage in = garbage out*' problem familiar to all informaticists. Much of the value of that data will remain unrealized, as the cost to understand it will continue to increase with the growth.

Data governance programs come in many different shapes and sizes, as they must be customized to address many different organizational factors. One way to create an implementation roadmap is to frame data governance design and implementation in the widely accepted people – process(es) – technology framework (The MDM Institute 2016). By doing so, you can map all major data governance drivers and activities to at least one of the pillars of the framework, which is already well-known to those in information technology-heavy disciplines. In the upcoming paragraphs we will describe data governance in this manner.

6.5.1 People and Roles in Data Governance

One of the most difficult aspects of implementing a data governance program may be managing the people and roles of those involved in the different aspects of data management. Often there is the need to centralize, consolidate, or dramatically alter job descriptions and duties to meet the larger needs of the healthcare organization. This involves changing data activities in a way that may not initially seem advantageous to those affected. Data management optimization will usually require more active sponsorship from senior leaders than previously committed, and inevitably all levels of an organization undergoing anything other than a trivial data governance effort will need to contribute (Loshin 2009). Those playing the most active roles in DG usually reside in a typical hierarchy such as the one shown in Fig. 6.3.

The Steering Committee typically provides guidance and helps set the direction of data governance, serving as an advisory board and ratification body. The Data Governance Operational Workgroup(s) are the main tactical force, relaying the frontline data management needs and carrying out the approved data governance activities. The Data Governance Council acts as the connecting body between the Operational Workgroup and Steering Committee, working to communicate

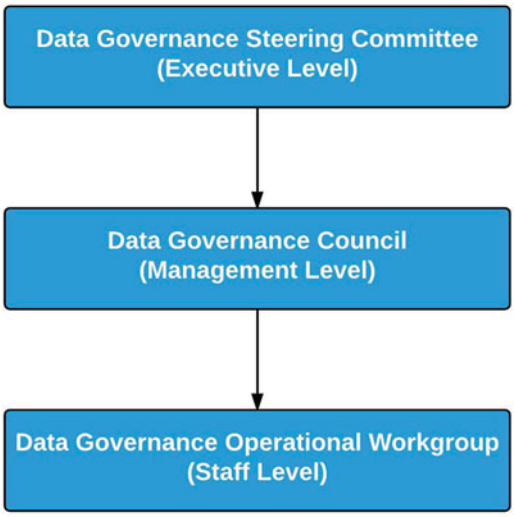


Fig. 6.3 Hierarchy of roles in a data governance program

Table 6.2 Description of roles in a data governance program

Data governance role	Description
Executive sponsor	Can be 1 person or an entire committee
Leader	Leads all data governance activities
Program manager	Acts as initiative manager, coordinator
Data domain owner	Data domain subject matter expert, works with Data Steward on all data governance activities related to their domain
Data steward	Carries out data governance work processes under direction of the Data Domain Owner and in alignment with data governance processes and policies
Enterprise data steward	Data Steward with a holistic view of the organization's data needs, coordinates work of various Data Stewards
Data analyst	Interacts with data output from data governance processes, to serve the needs of Users
User	Consumers of the data
Data architect	Informs the design of the data warehouse and data governance-related Information Technology
Data modeler	Builds the data models necessary to house the data
Database administrator	Carries out the day-to-day maintenance of the data governance-related Information Technology hardware
Security analyst	Helps set security policies for access, informs hardware build, monitors for intrusions and unapproved access
Business intelligence developer	Assists others in drawing insights and value from data

progress, problems, and proposed solutions across all groups. It is very common to have individuals working within multiple levels.

There are many different roles that can be associated with data governance programs, depending on the size and scope of the overall program. Table 6.2 lists some

example roles, but is by no means all-inclusive. Because many of these roles have overlapping skillsets (particularly the technical roles), it is common to create a Roles, Accountability, and Responsibility Matrix to delineate 'who does what' in the data governance program.

6.5.2 Data Governance Processes

The foundation for any successful data governance program is designing and implementing the activities the various roles will carry out (Haines 2015). Standardizing workflows and acceptable practices will provide optimal expectation management, efficiency, reproducibility, and most importantly, reliability of the data governance program's output. A few of the more salient activities data governance programs focus on are highlighted here (Loshin 2009). These processes assume the data has been modeled and loaded into a data warehouse or other data source.

6.5.2.1 Training/Onboarding

After roles are established and personnel are assigned, the onboarding process should begin. During this time personnel are oriented to the data governance program, trained to fill skillset gaps, and coached through their first data governance activities by the Enterprise Data Steward or other Trainer. Training often involves didactics, independent learning, or side-by-side, 'at the elbow' interactive training.

6.5.2.2 Policies

Documentation in the form of policies helps guide all users to acceptable practices. They are a source of reference to settle disputes and guide expected actions, among other things. Policies should be linked and integrated to already existing organizational documentation so as to avoid duplication, conflicting information, or increased maintenance. Typical data governance policy subjects include data security, access, ownership, data quality, and retention, just to name a few.

6.5.2.3 Data Quality

Data quality can be divided into many different dimensions. Common dimensions include integrity, completeness, consistency, accuracy, timeliness, conformity, validity, duplication, range, and accessibility. Each data governance program should determine the dimensions that make the most sense for their situation.

Data quality begins with an assessment of the data and its characteristics. This initial activity involves taking a high-level view of the data and documenting the

findings, an activity known as **data discovery and profiling**. Simple evaluations of the data can give one insight into how successful the extract-transform-load (ETL) process was. The next step is to define the data using documentation of data entities and terms in a **business glossary**. This tool serves as a common catalog of the data for other users,, thereby increasing transparency and efficiency of access. **Business rules** are used to transform the data into useable, semantically appropriate datasets that can be used to answer questions or provide insight. Each of these manipulations or transformations must be documented as well, alongside the **data's lineage** so future troubleshooting of output is possible without much effort. Some of these processes can be automated through software solutions and the results can be shared through **monitoring metrics** and viewing **data dashboards**. Similar dashboards can be used to document the value of the data governance program, demonstrating and justifying the program through a return on investment (Informatica 2013).

6.5.2.4 Communication

An effective communication strategy should be constructed to convey the intent, value, timeline, and impact of the proposed data governance program. Large organizations will find that many different communication formats and channels are required to reach their many data constituents, and that different variations of the message may need to be crafted for each group. Messages should be succinct but contain enough information to be useful for the receiver. It is recommended that for larger data governance programs a repository of materials be made available for interested parties to view. References to these centrally located materials will enable the primary message to be short and more likely to be read. Data governance communication leaders should also take advantage of already established communication channels when possible.

6.5.3 Data Governance Technology

Data governance technology, like most other aspects discussed in this chapter, is very diverse and dependent upon the data environment of the healthcare organization. Data governance can be applied to any information technology system that generates, stores, or otherwise processes data. Generally speaking, however, healthcare institutions are large and complex both in physical structure and technological infrastructure. The larger and more robust the data environment is, the more difficult and complex the technologic needs for data governance tend to be. Solutions may range from application of new data governance policies and processes across existing data tools to implementation of entire suites of software, hardware, and accompanying policies, procedures, and personnel restructuring. An exhaustive list of all types of data governance technology is out of scope for this chapter, but we will briefly describe tools and solutions in the next few paragraphs as an introduction.

Most institutions that start a robust data governance program have also purchased hardware that allows them to build centralized data repositories. The hardware is often engineered to provide the structure for an Enterprise Data Warehouse. Enterprise Data Warehouses usually accept data feeds from multiple distributed data sources and act as the cross-functional platform for integrating data in a central location. Other related terms the reader may come across include data farms, data universes, data lakes, or central data repositories. Each of these is closely related, but structured differently to enable different functions. Data governance should strongly align with the capabilities of the program's technologic infrastructure, as well as the capabilities and expected functions.

There are many different data governance software solutions currently on the market, including many modular options from vendors. The software selected should match priorities and fulfill the needs of the institution. Products are available to help with every stage of the implementation, from hardware management, data modeling, to data quality and monitoring. More advanced activities supported include Master Data Management and Metadata Management, which are often implemented at the more advanced and mature stages of data governance programs (Informatica 2013). Most of these tools allow users to work collaboratively to enforce and provide data governance best practices.

Even more numerous are the tools available for presenting data to users. Many of these tools are meant to enable a self-service model once the underlying data is of high enough quality to be made available, which helps to relieve data acquisition bottlenecks from dependencies on report writers, data analysts, and stewards. There is no shortage of business intelligence and data visualization software options on the market today.

6.6 Conclusion

The use of integrated data repositories can enable the efficient use of electronic health record data for research purposes. Care should be taken, however, in choosing a data model. Each model was designed for a specific purpose, and as such, has been optimized to meet the needs of that particular use case. As a result, there are tradeoffs when trying to use a model for a purpose for which it was not intended. Procedures now exist that allow users to transform data from one model to the other, but one cannot assume 100 % fidelity in a transformation. When looking at the use of data across an institution, data governance is crucial to ensuring that decisions are coherent across the enterprise. And while there are many strategies that can be used to implement data governance, whatever hardware and software is ultimately selected, implemented, and used, the old informatics adage of incorporating the technology into workflow still holds true – the people, processes, and technology must all mesh seamlessly and without friction for any one part to function at the highest level.

Acknowledgements This publication describes work that was funded in part by NIH/NCATS (UL1TR000017 / UL1RR026314, UL1TR001425). Its contents are solely the responsibility of the authors and do not necessarily reflect the views or opinions of the NIH.

References

Blumenthal D. Stimulating the adoption of health information technology. N Engl J Med. 2009;360(15):1477–9.

Blumenthal D. Launching HITECH. N Engl J Med. 2010;362(5):382–5.

Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med. 2010;363(6):501–4.

Califf RM. The Patient-Centered Outcomes Research Network: a national infrastructure for comparative effectiveness research. N C Med J. 2014;75(3):204–10.

Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, Raebel MA, Beaulieu NU, Rosofsky R, Woodworth TS, Brown JS. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. Pharmacoepidemiol Drug Saf. 2012;21 Suppl 1:23–31.

Dhir R, Patel AA, Winters S, Bisceglia M, Swanson D, Aarnodt R, Becich MJ. A multidisciplinary approach to honest broker services for tissue banks and clinical data – a pragmatic and practical model. Cancer. 2008;113(7):1705–15.

Eggebraaten TJ, Tenner JW, Dubbels JC. A health-care data model based on the HL7 reference information model. IBM Syst J. 2007;46(1):5–18.

Haines R. What is "just-enough" governance for the data lake? 2015. Retrieved March 11, 2016, from https://infocus.emc.com/rachel_haines/just-enough-governance-data-lake/.

Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, van der Lei J, Pratt N, Noren GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–8.

Informatica. Metadata management for holistic data governance. 2013.

Kimball R. The data warehouse toolkit: practical techniques for building dimensional data warehouses. New York: Wiley; 1996.

Kimball R. The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses. New York: Wiley; 1998.

Kimball R, Caserta J. The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data. IN, Wiley: Indianapolis; 2004.

Kimball R, Ross M. The data warehouse toolkit: the complete guide to dimensional modeling. New York: Wiley; 2002.

Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc. 2012;19(2):181–5.

Levenson D. Sweeping changes proposed for "Common rule": update would require mandatory informed consent for use of human biospecimens in research. Am J Med Genet Part A. 2015;167(12):viii–ix.

Loshin D. Master data management. Amsterdam/Boston: Elsevier/Morgan Kaufmann; 2009.

Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE – an integrated standards-based translational research informatics platform. AMIA Annu Symp Proc. 2009;2009:391–5.

Mackenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc. 2012;19:e119–24.

Marsolo K, Corsmo J, Barnes MG, Pollick C, Chalfin J, Nix J, Smith C, Ganta R. Challenges in creating an opt-in biobank with a registrar-based consent process and a commercial EHR. J Am Med Inform Assoc. 2012;19(6):1115–8.

Mendis M, Wattanasin N, Kuttan R, Pan W, Philips L, Hackett K, Gainer V, Chueh HC, Murphy S. Integration of Hive and cell software in the i2b2 architecture. AMIA Annu Symp Proc. 2007;1048.

Mendis M, Phillips LC, Kuttan R, Pan W, Gainer V, Kohane I, Murphy SN. Integrating outside modules into the i2b2 architecture. AMIA Annu Symp Proc. 2008;1054.

Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc. 2006;1040.

Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124–30.

Murphy SN, Gainer V, Mendis M, Churchill S, Kohane I. Strategies for maintaining patient privacy in i2b2. J Am Med Inform Assoc. 2011;18 Suppl 1:i103–8.

Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, Smith B. The national center for biomedical ontology. J Am Med Inform Assoc. 2012;19(2):190–5.

Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, Morris JA. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc JAMIA. 2010;17(6):652–62.

Rijnbeek PR. Converting to a common data model: what is lost in translation?: Commentary on "fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model". Drug Saf. 2014;37(11):893–6.

Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G, Steiner JF. The HMO research network virtual data warehouse: a public data model to support collaboration. EGEMS (Wash DC). 2014;2(1):1049.

Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, Sim I, Chute CG, Solbrig H, Storey MA, Smith B, Day-Richter J, Noy NF, Musen MA. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS. 2006;10(2):185–98.

SACHRP. Attachment A to January 24, 2011 SACHRP Letter to the Secretary – FAQs, Terms and REcommendations on Informed Consent and Research Use of Biospecimens. 2010.

The MDM Institute. What is data governance. 2016. Retrieved March 11, 2016, 2016, from http://0046c64.netsolhost.com/whatIsDataGovernance.html.

Title 45 Code of Federal Regulations. "Title 45 Code of Federal Regulations Parts 160, 162, 164: Combined Regulation Text." from http://www.hhs.gov/ocr/privacy/hipaa/administrative/combined/index.html.

Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? Ann Intern Med. 2009;151(5):359–60.

Xu Y, Zhou X, Suehs BT, Hartzema AG, Kahn MG, Moride Y, Sauer BC, Liu Q, Moll K, Pasquale MK, Nair VP, Bate A. A comparative assessment of observational medical outcomes partnership and mini-sentinel common data models and analytics: implications for active drug safety surveillance. Drug Saf. 2015;38(8):749–65.

Zhang GQ, Siegler T, Saxman P, Sandberg N, Mueller R, Johnson N, Hunscher D, Arabandi S. VISAGE: a query interface for clinical research. AMIA Summits Transl Sci Proc. 2010;2010:76–80.

Chapter 7
Laboratory Medicine and Biorepositories

Paul E. Steele, John A. Lynch, Jeremy J. Corsmo, David P. Witte, John B. Harley, and Beth L. Cobb

Abstract Biorepositories provide access to specimens for biomarker investigation of subjects with or without a given condition or clinical outcome. They must record collection, transport, processing, and storage information to ensure that specimens are fit-for-use once a particular analyte has been identified as a candidate biomarker. Ongoing (post-collection) clinical and outcome documentation provides more value to researchers than a static, clinical snapshot at the time of collection. Frequently,

P.E. Steele, M.D. (✉)
Department of Pathology and Laboratory Medicine, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Ave, ML 1010, Cincinnati, OH 45229-3039, USA
e-mail: paul.steele@cchmc.org

J.A. Lynch, Ph.D.
Department of Communication, University of Cincinnati McMicken College of Arts and Sciences, Cincinnati, OH, USA
e-mail: john.lynch@uc.edu

J.J. Corsmo, M.P.H.
Cincinnati Children's Research Foundation, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
e-mail: jeremy.corsmo@cchmc.org

D.P. Witte, M.D.
Departments of Pediatrics and Pathology and Laboratory Medicine, University of Cincinnati College of Medicine, Cincinnati, OH, USA

Department of Pathology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
e-mail: david.witte@cchmc.org

J.B. Harley, M.D., Ph.D.
Department of Pediatrics Center for Autoimmune Genomics and Etiology (CAGE) and Cincinnati Biobank, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, OH, USA

US Department of Veterans Affairs Medical Center, Cincinnati, OH, USA
e-mail: john.harley@cchmc.org

B.L. Cobb, M.B.A.
Center for Autoimmune Genomics and Etiology (CAGE) and Cincinnati Biobank, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, ML 15012, Cincinnati, OH 45229-3039, USA

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_7

biorepository specimens are residua from those obtained for clinical management of a patient; whether routine clinical processing is acceptable, given the stability profile for a given analyte, will dictate whether non-standard processing will be required. Introducing nonstandard steps into clinical lab processing in order to preserve an analyte such as RNA or protein requires careful workflow planning. Accreditation for biorepositories is now available; standards have been developed to ensure that biorepository personnel, equipment, laboratory space, information systems, and policies/procedures, including those for quality management, meet the same high standards by which clinical laboratories are judged and accredited. An additional accreditation standard relates to the development of, and adherence to, policies surrounding informed consent. The current regulatory landscape for pediatric specimen research requires consideration of many issues around informed consent, assent, and reconsent at the age of majority for the collection and use of identifiable specimens for research. Consideration of these requirements based on the current (and evolving) regulatory landscape can be difficult, in light of pending legislative and regulatory changes. Issues surrounding return of incidental findings is another challenge for institutional review boards.

Keywords Biobanking of biospecimens • Consent and assent • Genomics • Incidental findings • Sample tracking

7.1 Introduction

The challenge in bioinformatics for the support of biorepositories is anticipating, and fulfilling, the data requirements of future biomedical researchers who wish to solicit samples that have been in storage, from days to decades.

What future questions will be posed to the biorepository manager? We can anticipate these questions:

- What is the specimen type, and how was it obtained?
- What were the demographics of the subject who was the source of the specimen?
- What were the clinical condition and the medical/family/social history of the subject?
- Since collection, how has the subject's clinical condition evolved?
- What were the environmental conditions, timing, and processing steps that characterized the pre-storage handling of the collected specimen, and how do these handling conditions relate to the stability of the analyte(s) of interest to the researcher?
- How has the specimen been stored, and how has the storage condition and time of storage affected the ability to recover the analyte(s) at levels equal to that at the time of collection?

- What is the nature of the consent obtained at the time of specimen collection with regard to research use, access to demographics, use of subject medical/family/social history before and after specimen collection, subject identification, and return of results (including incidental findings and genetic discoveries); what changes in this consent, if any, have occurred since the original consent?
- Can the above parameters be specified in an automated data search that could identify acceptable stored specimens for a given research investigation?
- Is the biorepository accredited, if so by what agency, and what operational implications does that accreditation carry?

7.2 Data Collection

Specimen type is of course a basic parameter for biobanks, but information on the method of collection, as well as the conditions under which collection occurred, must also be captured. For example, a clean-catch urine, catheterized urine, and a urine obtained by suprapubic aspiration are all different with respect to probable microbial contamination. Liver tissue obtained during a surgical operation versus an autopsy may be vastly different in regards to the particular analytes present. For autopsies, the interval between the time of death and the collection of an autopsy specimen must be recorded, as must the temperature conditions of body storage, since analytes can degenerate or move variably from one body compartment to another after death. All of these factors must be recorded and available to maximize the specimen's value for research.

Alongside specimen data, basic demographic information must be effectively captured. While the biobank may have access to complete information on the subject, including birthdate and other protected health information (PHI), restrictions to the release of PHI in most situations and many jurisdictions would mean, for example, that instead of the birthdate, age at the time of collection would accompany the specimen. Since date of collection is also provided to researchers, the biobank must ensure that the specification of the age (e.g., in days, weeks, or months) does not run afoul of PHI definitions.

Basic demographic information includes the sex of the subject. While assignment of sex is typically straightforward for most subjects, it is more complex for transgender individuals, because the relation between the time of specimen collection and interventions (such as hormone treatment) may be relevant. Ancestry and ethnicity are of obvious interest; the accuracy with which ancestry and ethnicity have been determined may be uncertain. It is helpful to record capture method for said data.

These are the most fundamental demographics, but there are others that are relevant for particular research studies. For example, where the subject lives, and has lived, their current and past socioeconomic status, and other such data may be relevant (e.g., studies of environmental exposure, etc.). For other studies, body-mass index may also be an important variable to consider.

The clinical condition of the subject is of prime importance; the International Classification of Disease (ICD) codes are important data to capture and are often used to dramatically narrow a search of study relevant subjects, increasing the usefulness of biobanked materials. While helpful, the limited specificity of ICD codes demands further data interrogation. Knowing that a patient has been diagnosed with type I diabetes mellitus raises questions about the length of their illness, the status of comorbidities, type, length, and effectiveness of treatment, etc., even in the absence of the confidence that this diagnosis is the most appropriate for the patient in question. Their disease age-of-onset, past medical history, social history, family history, habits (e.g., smoking), are all likely to be important, as research studies involving biorepository samples will often be case-control studies for which the investigator will match study subjects and controls for many of parameters that can be found in the patient's history.

The value of a biorepository is enhanced by continuing to collect and store data on the clinical condition of subjects from whom samples have been collected. Not only could future clinical information include a disease that was present, but not yet diagnosed at the time of specimen collection, but also the record may also provide follow up information such as treatment response or clinical course that would be essential to a researcher who is studying a prospective biomarker for its value in treatment choice and/or assessment of prognosis.

Data documented at time of specimen collection may not include all of the demographic and clinical status data that might be desirable to a future researcher. The logical choices for accessing future data include medical record review, and if permitted, recontact of the subject. Both of these choices demand careful attention to the details of the relevant Institutional Regulatory Board-approved, protocol and consent. Both the protocol and the consent documents must specifically authorize researcher access to PHI (protected health information) and or subject recontact. Permission for direct contact with the subject may be prohibited, but in instances where the subject has approved of it, direct contact with the subject may permit an opportunity to gain information that does not reside in the medical record.

One effective strategy to deal with the need to protect the identity of research subjects, while permitting researchers to gain access to medical record information that was in the record at the time of specimen collection as well as information that was subsequently placed there, is the use of an "honest broker" (Choi et al. 2015). The broker stands between the medical record of the institution and the researcher; the broker can examine the nature of the consent and provide permitted data about the subject without revealing the identity of the subject to the researcher.

7.3 Specimen Handling, Processing, Labelling, and Storage

While many data elements can be retrieved from the medical record, details about the handling, processing, labelling, and storage of the specimen will not be in the medical record and so must be captured in the biorepository's software application.

There are a number of commercially-available software packages that may be used to capture this data (Boutin et al. 2016; McIntosh et al. 2015); in-house developed software may provide needed customized features, although the requirements demanded of the software by biobank accrediting standards make in-house-developed software a challenging, but not insurmountable choice (College of American Pathologists 2015).

The data elements relevant to handling, processing, labelling, and storage may be found in Table 7.1.

A common unique identifier is the patient medical record number; if the patient sample comes from a multi-institutional setting such as a hospital consortium, there may be a master index number that supercedes the medical record number. The medical record number is unique to the subject, but is typically used for all encounters, and so an encounter-specific, unique identifier is also needed; typically, this encounter-specific number is the financial information number (FIN) or a clinical

Table 7.1 Data records for specimen handling, processing, labelling, and storage

Date, time, name and description of each sample handoff transaction
Medical Record Number
Unique identifier for each subject, including encounter identification
Project identifier and project name
Time and date of collection
Specimen type, including anticoagulant (tube type), preservative, additives
Specimen volume
Specimen concentration
Sample processing instructions including Standard Operating Procedure (SOP) name and version
Temperature or any other exceptional conditions during handling and processing, if outside the SOP
Conditions of initial centrifugation and any subsequent centrifugation for derivative samples
Method used for preparing derivative samples
Quality and purity of parent and derivative samples
Time and date of entry into storage
Unique identifier for each aliquot of each parent specimen (and each of its derivatives)
Number of aliquots, with volume of each
Date and time of each freeze and thaw with the residual volume
Storage location: room, freezer, shelf, rack, box, position within box
Storage barcode identifier
Sample requester name and contact information
Sample requester designee information
Sample request IRB protocol ID
Total # of samples retrieved per retrieval
Date of sample retrieval request
Date of sample retrieval provision
Sample request description of each sample (concentration, Optical Density, container type)
Sample request release information: signature of releasee, date and time

laboratory intake identifier, often known as the accession number. If the subject was seen as part of a clinical trial, there is also typically a study encounter number.

The time and date of collection are routinely collected, but care must be taken to ensure that the time of collection is not defaulted as the time the specimen arrived in the laboratory; this default may be used when the data field for collection time is not a required field to be completed by the specimen collector.

"Plasma" is not a suitable description of a biological sample, as there are many varieties, including EDTA, lithium heparin, sodium heparin, and sodium citrate. Tube type is a preferred description, as it will contain additional information such as presence or absence of a separator gel, or presence of additives such as protease or RNAse inhibitors.

Specimen volume, as well as the volume of the final aliquot(s) is important when requests for distribution are processed.

Temperature conditions carry implications for stability that vary across multiple analytes (Ellervik and Vaught 2015; Riondino et al. 2015). A general consensus is that the more efficiently a sample can be processed and then stored at −80 degrees Centigrade (−150 degrees Centigrade, or liquid nitrogen for cells), the better. However, in much of biobanking, and especially in pediatric biobanking, the specimens collected are residua from specimens submitted for clinical testing. They are usually collected at room temperature, processed within minutes for specimens collected near a laboratory but not processed for hours for specimens collected in a remote location, held at room temperature during the testing process, then refrigerated for up to 1 week before being discarded (or directed to the biorepository). These conditions are suitable for genomic DNA extraction from anticoagulated blood or the subsequent evaluation of antibody titers in serum, but problematic for many other analytes. Cooperation between the clinical laboratory and the biorepository can help to overcome this challenge for unstable analytes, e.g., by parallel processing of a specimen that is split for clinical and research uses. Centrifugation conditions (rpm, *g*-force, temperature) during the initial separation of plasma or serum from cells, and during any subsequent centrifugation to produce derivative products, such as washed cells, must be recorded. The method for producing such derivatives, such as density gradient centrifugation, must be recorded.

The quality of the stored products should be recorded. The methods to rate quality are evolving but many have been proposed for DNA, RNA, and protein (Shabihkhani et al. 2014; Betsou et al. 2013). These quality measures can be obtained at the time of specimen entry into the biorepository, but they can also be applied to a sampling of specimens after various time intervals in freezer storage.

The time and date of entry into the storage system and all instances of removal from the storage system must be documented. Re-entry into storage following thawing, aliquoting a distribution sample, and re-freezing all need to be documented.

The labels used in storage must remain adherent to the tube; the use of bar-coded freezer tubes with no labels provides a suitable alternative. Each tube must be uniquely identified, down to the individual aliquots. The volume of sample within each tube must be recorded, and the recorded volume adjusted if the entire volume is not distributed.

Finally, the exact location of each tube (freezer room, freezer, shelf, rack, box and position within the box) is essential to facilitate quick retrieval; quick retrieval protects the contents of the freezer from excessive exposure to room temperature. Although robotic systems permit retrieval of a sample without exposing all the contents to room temperature when a sample is removed, they are expensive, especially when calculated for the amortized expense for each retrieved specimen.

7.4 Clinical Lab Processing Versus Biorepository Processing

Increasingly, clinical laboratories are pressured to turn around the processing, analysis, and result reporting of clinical laboratory specimens in order to speed the flow of patients through diagnostic and therapeutic episodes of care. Standardization with the intent to preserve uniformity and consistency in laboratory result values is the goal, with standard operating procedures written and strictly followed to ensure that specimens are handled the same way, every time.

The introduction of non-standard steps and non-standard documentation requirements into a high-volume, rapid throughout clinical lab processing area, for a small percentage of processed specimens that are destined for research use, lengthens turnaround times for all results and introduces risk that protocol-required steps for the research samples are omitted or performed with unacceptable delay. Further, the documentation of exceptional processing is often lax, adding to the difficulties thereby derived. Employees are stressed when put in a position to choose between processing a specimen from a sick child in the intensive care unit or risking a protocol violation with a research sample that must be processed within a tight time frame. Processing steps for research specimens, including those going into the biorepository, often involve producing multiple aliquots, each of which must be labelled and frozen quickly; this process is unlike that performed for clinical laboratory samples, and it is time-consuming.

Furthermore, documentation suitable for clinical lab processing is captured, often automatically, by clinical laboratory information systems, but the data capture points are insufficient for research or biorepository requirements. Table 7.2 lists the typical data elements captured by clinical laboratory information systems, and Table 7.3 lists the additional data elements that research protocols and/or

Table 7.2 Data elements captured by clinical laboratory information systems

Time and date of collection
Time and date of laboratory receipt
Time and date of analysis
Time and date of preliminary result acceptance
Time and date of final result verification

Table 7.3 Data elements often required for research but not captured by clinical laboratory information systems

Time of entry into the centrifuge
Time of centrifugation
Temperature of centrifugation
Centrifugation *g*-force
Model and serial number of centrifuge
Time of exit out of the centrifuge
Time of entry into freezer or refrigerator storage

biorepositories typically require that are not captured as a part of routine clinical lab processing; capturing the latter data may require a research coordinator accompanying the sample to the clinical lab, either personally performing the processing steps or at least recording the times and other data.

A more satisfactory approach is to separate clinical lab processing from research processing, creating a fast lane and slow lane, respectively. This approach is admittedly resource-intensive, as samples for the slow lane may come at any time on any day, and the volume of samples in the slow lane may seem insufficient to justify 24×7 staffing.

7.5 Accreditation

The College of American Pathologists (CAP), which accredits clinical laboratories around the world, has recently begun to accredit biobanks (College of American Pathologists 2015). The CAP requirements for biorepository informatic technology (IT) systems mirror their requirements for the mission-critical IT systems for clinical laboratories.

Of note, the information systems must be robust, i.e., automatically backed-up and able to be restored quickly when a hardware or software failure occurs. Logon-associated security levels, tied to responsibility and authority, must reflect roles that are certified by the institution. Logons must be secure and not shared. An audit capability must track additions, deletions, and corrections in the system, positively identifying which staff member made any such changes. Staff must include a senior administrator(s) who is responsible for approving all changes to the system. Standard operating procedures must direct staff activities when downtimes occur, and they must specify conditions (such as checks on data integrity) that are required before re-starting the system after a downtime. There must be documentation of the software functionality including any custom alterations to the standard programs. Training records for staff must record their ability to successfully interact with the software program, and staff must have a defined escalation strategy when problems cannot be resolved in a timely manner.

Each specimen must be labeled with a unique identifier, and this identifier must be used as the specimen and aliquots/derivatives associated with it, move through the testing process. Labeling specimens in intermediate stages of processing as 1, 2, 3 or A, B, C, etc., is not acceptable. Time points for collection, processing, and storage must be captured, either automatically, or if necessary, manually. The storage records must indicate the storage temperature; ideally, the storage temperature is tracked continuously with frequent, e.g., hourly, temperature data points captured and recorded. Software for maintaining storage records should also track all additions to, and distributions from, the biorepository, providing information about who received a sample, what samples were provided, and when requests were processed and distributed. If there are parent-child relationships between specimens, the derivative specimens must be labeled so that those relationships are encoded. There must be a method to generate new labels as needed. Coding for sample identification, e.g., a letter code for different types of specimens, must be defined. All patient-related data associated with the specimens must be secured, and unauthorized access through programs or interfaces to this secure data must be prevented.

7.6 Ethical and Regulatory Issues

Before a biobank begins collecting or otherwise receiving samples it is important to consider several important ethical issues. An important factor in these considerations is the institution's decision around the type of samples and method of acquisition that will be used to collect samples for the biobank. An institution may begin with a strategic commitment to systematically collect and store some or all residual clinical samples for future unspecified research. For some institutions residual clinical samples may be a valuable source of research samples; in other cases an institution may choose to target specific types of samples, specific diseases/conditions or specific types of patients. Each of these decisions has its own set of unique cost, resource, logistic and scientific drivers. One could reason that the collection of residual clinical samples minimizes the impact to the patient providing the sample, e.g., limiting blood loss and invasive specimen collection procedures, and may have lower associated sample acquisition costs. However, broad-based consenting, often employed for institution-wide residual clinical sample collection, may present specific ethical challenges. Without careful distribution planning prior to biobank conception, this approach is vulnerable to resulting in a substantive number of samples being stored in the biobank and underutilized for scientific research. With that said, in this section we will explore several of the regulatory and ethical challenges associated with approaches to informed consent and return of secondary and incidental findings associated with a biobank designed to collect residual clinical samples.

7.6.1 Consent and Assent for Use of Residual Clinical Samples

One of most significant and complex decisions to be made at the outset of a biobanking initiative, focused on the collection of residual clinical samples, is whether and how to obtain informed consent and how to address the issue of informed assent. This choice will have implications for a hospital's electronic health record (or similar patient tracking) system, human resources, infrastructure, and patient flow. Once an institution gets beyond the basic requirement of choosing a consenting process and developing a document that complies with baseline federal, state, and local laws and regulations, more complex and strategic considerations will be needed. Specifically, in developing the consenting documents, the institution should give significant consideration to intended future use of the samples. For example, the institution needs to address questions such as whether the primary use of the samples will be for procedures such as large scale genomic sequencing (e.g. whole genome/whole exome sequencing). Secondly, the distribution rules are equally critical to patients and the users of the stored samples. Patients and families may have strong feelings, both positive and negative related to the sharing of their sample with industry. Therefore, it is important that the institution address whether the samples will be for internal use only, made available to external academic collaborators, provided to external industry partners and/or potentially shared with international collaborators. Once questions around future sample use are addressed, the institution must also address whether and how available clinical and phenotypic information will be linked to and available for use with the stored samples. The availability of this information and specifically its breadth and depth is vital to the utility of the stored samples for future research purposes. From a strategic standpoint, institutions must decide whether to store consent documents (and in most cases a HIPAA (Health Insurance Portability and Accountability Act) authorization) in the medical record, thereby creating a link to all available clinical information. Alternatively, institutions may elect to obtain permission at the time of consent to recontact the patient in the future if there is a need to link their stored sample to their available clinical information.

Both of these decisions present unique challenges, including associated upfront investments in time and resources or increased burden at the time of sample use. Some of these challenges may be addressed at the time of sample distribution; however, most important is that the consenting documents used to allow the acquisition of samples into the biobank completely address each one of the institutional decisions related to the planned use, desired strategy for sharing, and plans for linking data to the stored samples. This will ensure that the patients and families provide an adequate consent to cover the desired future uses of the samples. Creating and maintaining a biobank requires a significant investment, and learning years down the road, after tens of thousands of people have been consented, that the consents obtained for the samples are not adequate for the desired uses of the samples is a disastrous error that could mark the failure of a biobank. However, this vulnerability can usually be easily mitigated with sufficient and robust advanced planning.

Within the context of a pediatric setting, biobank managers may find a significant value added from engaging both parents and children in the planning process, particularly regarding the development of consent documents and processes as the interests of the biobank, the research community and parents and children are not necessarily the same (Avard et al. 2011). From an IRB oversight perspective the literature generally supports the notion that the consent/permission of one parent is typically sufficient (Brothers et al. 2014). In the spirit of meeting the ethical underpinnings of the Belmont Report, institutions should also consider the need to provide developmentally-appropriate materials to explain the biobank. Importantly these documents should address the questions of how, why and by whom would the minor's stored samples be used in the future. Although there is some variability in the literature on this topic, a reasonable approach would be to consider directly engaging the great majority of children in a formal assenting process around the ages of 10 or 11 (Brothers et al. 2014). Lastly, with regard to pediatrics, and of specific importance to informaticians charged with developing and maintaining systems to support biobank operations, if the biobank will retain any ability to identify a specific sample, donor systems should support the tracking of patient age such that the biobank can be alerted when a patient reaches the age of majority (typically 18 years of age), triggering processes to secure the independent adult consent of the patient for the continued use of the stored samples and/or for use of future samples collected and corresponding linked clinical data.

Once decisions driving the content of the consenting documents have been made by institutions, typically in collaborations with their IRBs, there is a need to address the method that will be used, at least initially, to obtain informed consent. Several options regarding how the research informed consent will be obtained and documented should be considered: (1) request IRB approval for a waiver of the requirement to obtain an informed consent; however, proposals from US Department of Health and Human Services (HHS) and others, in the Notice of Proposed Rule Making for Changes to the Common Rule (80 FR 53931 September 8, 2015) suggest that this option may be prohibited in future revisions to the federal regulations; (2) incorporate the biobank-related consenting language into the institution's standard consent for medical treatment; (3) similar to #2, require an additional affirmative acknowledgement regarding participation in the biobank, either as a proactive opt in or as an opt out, as part of the standard consent for medical treatment; (4) develop a specific standalone consenting document for the biobank but incorporate the execution of that consent into the institutional clinic registration process; or (5) similar to #4, execute the consent as a separate research consenting process by dedicated personnel. Each of these options presents its own unique risk and benefit profile. Among the proposed changes, in September 2015, release of the Notice of Proposed Rulemaking for Changes to the Common Rule would be a requirement for written consent for research involving any biospecimens. Interestingly, the proposed rule change is specific to biospecimens only and does not introduce a comparable new consenting requirement to access personal information. Public comment has overwhelmingly opposed this proposal for numerous reasons, including its potential impact on the ongoing operations of *existing* biobanks (Cadigan et al.

2015). Regardless, institutions developing biobanks should anticipate the need for robust consent and assent processes, driven by potential federal regulatory changes as well as by a growing expectation for increased and proactive patient/stakeholder engagement.

In support of the growing national consensus of the importance of patient and other stakeholder engagement, and given the complexity and sensitivity of the issues, institutions may benefit from seeking input from patients, families and local community representatives regarding their perception of biobanking, types of specimen sharing, linkage to medical record data and the consenting process. Although the opinions of patients and families may be colored by the realities of the diseases and conditions that afflict them, and although local community representatives will likely approach these issues from their own internal biases, remarkably consistent commentary has been obtained from these groups of stakeholders in our own experience which ultimately was utilized in the structure and implementation of our local biobank.

Although we leveraged considerable input from patient, family and community stakeholders we also made several internal strategic decisions regarding the overall purpose of our local biobank. Our decisions have resulted in some successes, such as a custom-built informatics interface with our electronic health record. This solution allows for easy presentation and documentation of a patient or guardian biobank consent decision during the regulatory clinic visit registration workflow. In our instance of the consent process, a variety of consent responses are available for participants: (1) Agreement to participate with a desire for return of incidental findings; (2) Agreement to participate with a prohibition on the return of incidental findings; (3) Consent deferred or undecided; (4) Decline to participate. The decision of the patient or guardian is recorded in the electronic medical record system, which is then tracked with any potential biobank's sample in the acquisition and management system. Consistent with local IRB approval requirements, patients over age 10 are engaged in a verbal assenting process, with no specific requirement to document the assent decision.

With regard to the design of an informatics platform it is important to clearly distinguish "consent deferred" from "decline to participate". "Consent deferred" indicates that the patient/family is unable to make a decision or that the registration team decides based on clinic flow/volume that there is not time to sufficiently present the biobank consent at the time of registration. This option is anticipated to be temporary and will terminate when the patient or guardian either obtains additional information or presents at another clinic with sufficient time to consider participation in the biobank. In contrast, "decline participation" is a definitive choice of the patient or guardian regarding the use of their residual clinical samples as part of the biospecimen repository. This is intended to be a more permanent decision with the consenting system holding that refusal for 1 year following the decision, at which point the patient again becomes eligible to be approached regarding study participation. Finally, research participants must always be able to withdraw from a study. Therefore, an informatics platform supporting a biobank must be able to capture a withdrawal of consent, inclusive of its effective date, to ensure no confusion when samples are considered for collection or release (see Sample Retention).

More broadly, informatics input into storing and propagating consent data is essential. At the most basic level, researchers must track consent information corresponding to each study sample. Storage and tracking capacity of said data is not usually offered or managed by sample tracking systems. With regard to biobanks that will store samples collected from children, it may also be important to store information like the child's date of birth, as well as information about the person that provides the parental consent (e.g. name and relationship). This information may be needed for verification, especially if questions about the validity of the parental consent are raised at a later date. For an institutional biobanking effort, this is a daunting challenge and can only be accomplished using an informatics solution. Additionally, although there is some debate around how much effort institutions should expend in contacting and obtaining consent from a research participant once they reach the age of majority, generally biobanks should be prepared to engage in some level of effort to obtain a typical research consent from minors once they reach the age of majority, unless samples in the biobank are stored in a completely anonymized manner (Brothers et al. 2016). Biobanks are encouraged to proactively engage with their local IRBs in developing and understanding the limits of these efforts, including the ability to continue using samples once attempts to secure adult consent have been exhausted.

7.6.2 Return of Incidental Findings

Other than issues related to how consent is obtained, the other most significant ethical challenge facing biobanks is how they will manage the discovery of secondary and incidental findings. These terms are sometimes used interchangeably; however, for our purposes, we will differentiate these terms. "Secondary findings" are those findings that are discovered as a result of proposed future research involving banked samples. In general, biobanks should expect that researchers planning such secondary research projects should incorporate plans for handling/reporting secondary findings as part of the research plan. In the case of this future research and potentially important secondary findings, biobank personnel need to ensure that the consents used to obtain and store the biobank samples allow the proposed future research, and that any institutional assurance will be followed.

Any discussion of biobanking in the post-genomic era would be incomplete without a meaningful discussion of return of 'incidental findings'. This group of findings represents information that is learned about an individual that is unexpected or otherwise goes beyond the scope of the planned research or clinical evaluation. At the close of the twentieth century, bioethicists believed return of research results should occur rarely, if at all (National Bioethics Advisory Commission 1999). Yet, this position was never universally accepted, and bioethicists have moved toward an ethical obligation to return incidental findings to research participants, grounded in the Belmont Report's principles of respect for persons, beneficence, and justice (Presidential Commission for the Study of Bioethical Issues 2013; Kohane et al. 2007; Wolf et al. 2008) Current guidelines for adult and pediatric biobanks suggest

that if incidental findings are to be returned, then the possibility should be raised when informed consent is obtained (Brothers et al. 2014; Presidential Commission for the Study of Bioethical Issues 2013; Jarvik et al. 2014). Some have argued that if biobanks *have* genetically based incidental findings on hand, they are ethically obligated to return them, but they are not obligated to *search* for incidental findings (Jarvik et al. 2014). Others have argued that certain incidental findings are so important, they must always be sought and returned (Green et al. 2013). Given the ongoing debate about what results should be returned, including whether or not certain results are even actionable in a pediatric settings, some have suggested it is acceptable for a biobank to choose whether they wish to return results obtained from pediatric samples (Brothers et al. 2014).

Implementing a strategy to meet a perceived ethical obligation to return incidental findings can be challenging and requires understanding values and preferences of the local community. We designed a consenting process with two levels of participation. One level (participation with return of incidental findings) allows the patient or guardian to receive information regarding individual, clinically actionable incidental findings discovered during future research with the stored samples. A second option (participation without return of incidental findings) allows samples to be included in the biobank with the understanding that there is no mechanism for returning incidental research findings to the patient or guardian. Consultation with our local community has greatly informed our strategy. As described, during the consenting procedure, the patient or guardian is asked if they would like incidental findings returned or not. This selection is recorded as a component of the consent document; however, this recorded decision only allows our institution to initiate a multiple step process to evaluate whether a particular incidental finding is eligible for return. We will consider several of the major components of this process in more detail below; however, in general if researchers believe that they have learned important incidental information about a particular sample the recommended procedure follows the algorithm in Fig. 7.1

In executing this algorithm the IRB is attempting to establish a framework for what results should be returned; who should return them and who should receive

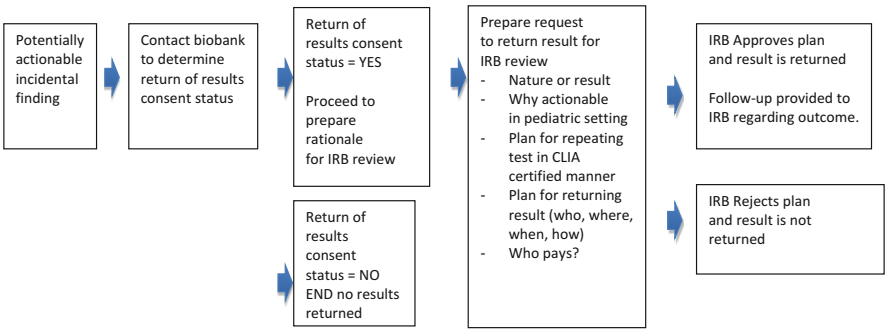


Fig. 7.1 Algorithm for incidental finding return

these results; how the results should be returned; and what additional obligations researchers and clinicians have to ensure participants have sufficient resources to act on the report of incidental findings (Wolf et al. 2008; Fernandez et al. 2003). All of these issues are further complicated in pediatric settings by the tripartite relationship among clinician, child, and parent and the evolving intellectual capacity of pediatric, especially adolescent, patients.

7.6.3 What Results Should Be Returned?

Bioethicists argue that results that can be actionable ought to be returned. "Actionability" covers three areas: (1) clinical utility, (2) reproductive planning and decision making, and (3) life-planning and decision-making (e.g., reporting *APOE4* which increases risk for Alzheimer's). While this standard seems intuitive, it is complicated by two factors. Research results are not always obtained with an FDA approved test and their significance is not always clear (Bookman et al. 2006; Miller et al. 2008; Shalowitz and Miller 2005). In order to be FDA approved in the United States, a test undergoes extensive testing and validation to ensure that the test is robust and results are both specific and sensitive. Tests that are not approved are of less certain validity. Also, many research laboratories are not CLIA certified to perform clinical testing (Wolf et al. 2008; Bookman et al. 2006; Clayton 2005). In the United States, clinicians and researchers are currently faced with an ethical conundrum brought on by narrowly interpreted CLIA regulations versus HIPAA privacy regulations. Specifically, while it makes sense to simply inform a patient and/or clinician of a non-validated research lab result that may need to be clinically validated in a CLIA-certified lab, doing so is interpreted by some as a CLIA violation that may invoke potential monetary penalties, since patients and clinicians may be tempted to use the results for clinical/diagnostic purposes. This concern is not supported by HIPAA regulations which suggest that the patient has a right to any information in possession of the healthcare entity that might be relevant to the individual's current or future health.

The landscape for evaluating incidental findings is further complicated in a pediatric setting when faced with an incidental finding about an adult-onset condition. IRB(s), ethicists and biobank professionals are faced with the decision of whether incidental findings for adult onset conditions (e.g., *BRCA1* and *BRCA2* mutations that increase risk for breast cancer) should be returned to a child or guardian based on the consent of the parent or guardian, or if the child should be given the right to consent to having their results returned once he/she reaches the age of 18. An American College of Medical Genetics (ACMG) policy statement suggested that large-scale biobanks and genomic studies should return results for 56 genes with clinical significance for life-long and adult-onset conditions (Green et al. 2013), but a joint ACMG-American Academy of Pediatrics policy statement advises that children should not be tested for adult onset conditions (American

Academy of Pediatrics 2013). While these statements do not directly address the issue of incidental findings in minors, they establish a framework for differentiating the burden for return of incidental findings in pediatric versus adult patients.

7.6.4 Who Should Return Results and Who Should Receive Them?

There is no consensus on who should return incidental findings from research. While researchers may be most familiar with the science supporting the results, they may not have the appropriate clinical experience or training and most likely lack the requisite genetic counseling experience to effectively explain results to participants (Avard et al. 2011; Wolf et al. 2008; Bookman et al. 2006; Sharp 2011) There is also concern about researchers contacting individuals with whom they have never had contact, e.g., as in research performed on stored samples (Wolf et al. 2008). On the other hand, primary healthcare providers will have a rapport with their patient and his or her parents or guardians, but the primary healthcare providers are not likely to have the expertise to interpret the results of genetic tests, especially those that are not frequently performed in a clinical setting. Typically in these circumstances the preferred approach is to have a qualified professional genetic counselor involved in the return of any unusual genetic result, including incidental findings. Finally, there is the challenge of providing results to providers and patients in an easily accessible format. The electronic health record has been identified as an ideal vehicle of returning information (Brothers et al. 2016); however, the EHR cannot be used when research results come from non-CLIA approved laboratories. In research-intensive institutions, a research patient data warehouse, research registries, and special software applications to connect researchers with patients may need to be developed (See Chap. 6, Data Governance and Strategies for Data Integration). As described previously, the role of the institutional IRB should not be overlooked in developing these processes to be certain they comply with applicable institutional policies, laws and regulations.

There is also some question about who should receive results. As discussed previously, it is fairly well established that participants in a biobank or similar research study should be given the opportunity either to request that available incidental findings be returned or to refuse return of such results (Wolf et al. 2008; Fernandez et al. 2003; Bookman et al. 2006; Clayton 2005). Yet, some argue that others, such as immediate family members, might benefit as well, in the event that serious health risks are identified (Avard et al. 2011; Black and McClellan 2011). Pediatric research raises additional complications since minors, including older adolescents, do not technically have the authority to make decisions related to return of results, and no clear guidance exists for managing the ethical issues raised when parents and older adolescents strongly disagree on whether a result should be returned (Avard et al. 2011; Wolf et al. 2008). While current regulations allow researchers to return

these results to parents or guardians, the ethical reservation remains that current and future autonomy is compromised when results are returned over the objection of an adolescent. Additionally, there is a concern that parents or guardians may make decisions regarding return of genetic results that do not represent any change in clinical risk during childhood. Patients are consented for enrollment in the biobank during registration at their initial visit to our hospital. As part of our institutional broad-based consent project, Better Outcomes for Children, parents or guardians of patients have been asked to provide consent for utilization of residual clinical samples; over 80% have agreed to participate and have asked to be informed of results that the institution believes are indicators of a major disease that may be prevented or treated or any findings that the researcher deems would affect the subject's medical care.

7.6.5 What Is Owed to the Research Subject?

A final question to be considered is what is owed to the subject. If subjects agree to enroll in a biobank, it is impossible to identify all future research that might occur with those samples. Neither research subjects, the researchers, or biobank staff can anticipate what information will be produced. Therefore, a plan must be developed to help participants understand and process information relevant to incidental findings. Yet, the extent of that help is hard to define. Grants that support research rarely (if ever) provide resources to counsel participants, and researchers do not have external financial resources to contribute. Specific guidelines about additional support are still being developed, although a consensus seems to be developing that the minimal requirements include referral to additional sources of relevant expertise or services such as genetic counseling (Wolf et al. 2008; Bookman et al. 2006; McCullough et al. 2015).

The development of this chapter was supported by U.S. HHS grant U01 HG008666 for the Electronic Medical Records and Genomics (eMERGE) Network (https://www.genome.gov/27540473).

References

American Academy of Pediatrics Committee on Bioethics. Committee on Genetics, and American College of Medical Genetics, and Genomics Social, Ethical, and Legal Issues Committee: ethical and policy issues in genetic testing and screening of children. Pediatrics. 2013;131(3):620–2.

Avard D, Sénécal K, Madadi P, Sinnett D. Pediatric research and the return of individual research results. J Law Med Ethics. 2011;39(4):593–604.

Betsou F, Gunter E, Clements J, DeSouza Y, Goddard KA, Guadagni F, Yan W, Skubitz A, Somiari S, Yeadon T, Chuaqui R. Identification of evidence-based biospecimen quality-control tools: a report of the International Society for Biological and Environmental Repositories (ISBER) Biospecimen Science Working Group. J Mol Diagn. 2013;15(1):3–16.

Black L, McClellan KA. Familial communication of research results: a need to know? J Law Med Ethics. 2011;39(4):605–13.

Bookman EB, Langehorne AA, Eckfeldt JH, Glass KC, Jarvik GP, Klag M, Koski G, Motulsky A, Wilfond B, Manolio TA, Fabsitz RR, Luepker RV. Reporting genetic results in research studies: summary and recommendations of an NHLBI working group. Am J Med Genet A. 2006;140(10):1033–40.

Boutin N, Holzbach A, Mahanta L, Aldama J, Cerretani X, Embree K, Leon I, Rathi N, Vickers M. The information technology infrastructure for the translational genomics core and the partners biobank at partners personalized medicine. J Pers Med. 2016;6(1):6.

Brothers KB, Lynch JA, Aufox SA, Connolly JJ, Gelb BD, Holm IA, Sanderson SC, McCormick JB, Williams JL, Wolf WA, Antommaria AH, Clayton EW. Practical guidance on informed consent for pediatric participants in a biorepository. Mayo Clin Proc. 2014;89(11):1471–80.

Brothers KB, Holm IA, Childerhose JE, Antommaria AH, Bernhardt BA, Clayton EW, Gelb BD, Joffe S, Lynch JA, McCormick JB, McCullough LB, Parsons DW, Sundaresan AS, Wolf WA, Yu JH, Wilfond BS, Pediatrics Workgroup of the Clinical Sequencing Exploratory Research (CSER) Consortium. When participants in genomic research grow up: contact and consent at the age of majority. J Pediatr. 2016;168:226–31.

Cadigan RJ, Nelson DK, Henderson GE, Nelson AG, Davis AM. Public comments on proposed regulatory reforms that would impact biospecimen research: the good, the bad, and the puzzling. IRB. 2015;37(5):1–10.

Choi HJ, Lee MJ, Choi CM, Lee J, Shin SY, Lyu Y, Park YR, Yoo S. Establishing the role of honest broker: bridging the gap between protecting personal health data and clinical research efficiency. Peer J. 2015;3:e1506.

Clayton EW. Informed consent and biobanks. J Law Med Ethics. 2005;33(1):15–21.

College of American Pathologists: Biorepository Checklist. 2015; [http://cap.org]

Ellervik C, Vaught J. Preanalytical variables affecting the integrity of human biospecimens in biobanking. Clin Chem. 2015;61(7):914–34.

Fernandez CV, Kodish E, Weijer C. Informing study participants of research results: an ethical imperative. IRB. 2003;25(3):12–9.

Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O'Daniel JM, Ormond KE, Rehm HL, Watson MS, Williams MS, Biesecker LG. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013;15(7):565–74.

Jarvik GP, Amendola LM, Berg JS, Brothers K, Clayton EW, Chung W, Evans BJ, Evans JP, Fullerton SM, Gallego CJ, Garrison NA, Gray SW, Holm IA, Kullo IJ, Lehmann LS, McCarty C, Prows CA, Rehm HL, Sharp RR, Salama J, Sanderson S, Van Driest SL, Williams MS, Wolf SM, Wolf WA, eMERGE Act-ROR Committee and CERC committee; CSER Act-ROR Working Group, Burke W. Return of genomic results to research participants: the floor, the ceiling, and the choices in between. Am J Hum Genet. 2014;94(6):818–26.

Kohane IS, Mandl KD, Taylor PL, Holm IA, Nigrin DJ, Kunkel LM. Reestablishing the researcher-patient compact. Science. 2007;316(5826):836–7.

McCullough LB, Brothers KB, Chung WK, Joffe S, Koenig BA, Wilfond B, Yu JH. Professionally responsible disclosure of genomic sequencing results in pediatric practice. Pediatrics. 2015;136(4):e974–82.

McIntosh LD, Sharma MK, Mulvihill D, Gupta S, Juehne A, George B, Khot SB, Kaushal A, Watson MA, Nagarajan R. caTissue Suite to OpenSpecimen: developing an extensible, open source, web-based biobanking management system. J Biomed Inform. 2015;57:456–64.

Miller FA, Christensen R, Giacomini M, Robert JS. Duty to disclose what? Querying the putative obligation to return research results to participants. J Med Ethics. 2008;34(3):210–3.

National Bioethics Advisory Commission: Research involving human biological materials: Ethical issues and policy guidance. 1999; vol. 1: http://bioethics.georgetown.edu/nbac/hbm.pdf.

Presidential Commission for the Study of Bioethical Issues. Anticipate and Communicate: Ethical management of incidental and secondary findings in the clinical, research, and direct-to-

consumer contexts. Washington, DC. 2013. http://bioethics.gov/sites/default/files/FINALAnticipateCommunicate_PCSBI_0.pdf.

Riondino S, Ferroni P, Spila A, Alessandroni J, D'Alessandro R, Formica V, Della-Morte D, Palmirotta R, Nanni U, Roselli M, Guadagni F. Ensuring sample quality for biomarker discovery studies – use of ICT tools to trace biosample life-cycle. Cancer Genomics Proteomics. 2015;12(6):291–9.

Shabihkhani M, Lucey GM, Wei B, Mareninov S, Lou JJ, Vinters HV, Singer EJ, Cloughesy TF, Yong WH. The procurement, storage, and quality assurance of frozen blood and tissue biospecimens in pathology, biorepository, and biobank settings. Clin Biochem. 2014;47(4–5):258–66.

Shalowitz DI, Miller FG. Disclosing individual results of clinical research: implications of respect for participants. JAMA. 2005;294(6):737–40.

Sharp RR. Downsizing genomic medicine: approaching the ethical complexity of whole-genome sequencing by starting small. Genet Med. 2011;13(3):191–4.

Wolf SM, Lawrenz FP, Nelson CA, Kahn JP, Cho MK, Clayton EW, Fletcher JG, Georgieff MK, Hammerschmidt D, Hudson K, Illes J, Kapur V, Keane MA, Koenig BA, Leroy BS, McFarland EG, Paradise J, Parker LS, Terry SF, Van Ness B, Wilfond BS. Managing incidental findings in human subjects research: analysis and recommendations. J Law Med Ethics. 2008;36(2):219–48.

Part II
Clinical Applications

Chapter 8
Informatics for Perinatal and Neonatal Research

Eric S. Hall

Abstract Effective biomedical informatics applications supporting newborn populations must go beyond simply adapting data systems or decision support tools designed for adult or even pediatric patient care. Within the neonatal intensive care unit (NICU), additional precision is required in the measurement of data elements such as age and weight where day-to-day changes may be clinically relevant. Data integration is also critical as vital information including the infant's gestational age and maternal medical history originate from the mother's medical chart or prenatal records. Access to these relevant data may be limited by barriers between institutions where care was provided, the transition between types of care providers (obstetrics to neonatology), appropriate privacy concerns, and the absence or unreliability of traditional identifiers used in linking records such as name and social security number. We explore challenges unique to the newborn population and review applications of biomedical informatics which have enhanced neonatal and perinatal care processes and enabled innovative research.

Keywords Maternal child health • Newborn • Perinatal • Record linkage

8.1 Infant Mortality and Neonatal Care

As many factors which contribute to infant death also influence the health of whole populations, the infant mortality rate (IMR) represents an important marker of population health (Reidpath and Allotey 2003). IMR is defined as the number of deaths among children under 1 year of age per 1000 live births during the same period of time (Blaxter 1981). Beginning in the 1860s, the collection of vital statistics data

E.S. Hall, Ph.D. (✉)
Departments of Pediatrics and Biomedical Informatics, Perinatal Institute and Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Ave, ML 7009, Cincinnati, OH 45229-3039, USA
e-mail: eric.hall@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_8

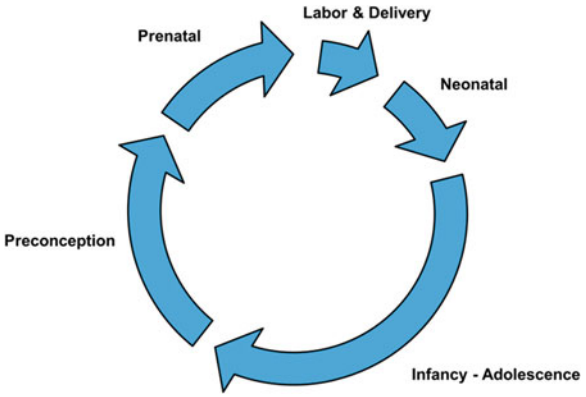


Fig. 8.1 Separation of lifespan data into lifestage-silos

motivated the British medical community to identify causes of their high infant mortality, which at the time was hovering around 450 per 1000 live births (Baines 1862).

Since then, vital records have played a key role in measuring the state of worldwide neonatal health. Spurred by major technological innovations, global infant mortality rates have declined dramatically over the past two centuries. By the 1990s, as a result of widespread adoption of neonatal intensive care units (NICUs) and other clinical innovations, infants born as early as 24 weeks of gestation had an approximately 50 % chance of survival – though neonatologists suspected they had reached the limits of viability (Allen et al. 1993). Corresponding with advances in neonatal care, all world regions have experienced a steady decline in infant mortality over the past several decades; though none are as dramatic as the 90 % reduction, which has occurred in the more developed regions of the world (United Nations 2011).

Although neonates are technically just young children, there are a number of important factors that differentiate the practice of neonatology from general pediatrics ranging from the typical care setting and workflow to the types of clinical problems that are managed. Along with the adult intensive care unit (ICU) or pediatric ICU (PICU), there is great potential for the NICU to benefit from computerization and decision support. While clinicians in all ICU settings may benefit from computer aided monitoring, data assimilation, and tools that aid and promote patient safety, fundamental differences in the NICU require special consideration.

Further, neonatologists strive not only to reduce infant mortality but to achieve long-term survival and healthy development among newborns, including those born preterm or having other significant complications (Saigal and Doyle 2008). To advance the overall state of neonatal health requires a broadened focus beyond care delivery during the neonatal period. Newborns are more likely to be healthy subsequent to a healthy, full-term pregnancy and healthier pregnancies are more likely to be achieved among women with a high level of health prior to pregnancy. Thus, to attain higher levels of newborn health requires an expanded focus on prenatal care, but also on preconception health as well as on the health of whole populations across the entire lifespan (Fig. 8.1). Advancements in data integration across the

lifespan are needed better understand the complex interactions among genetic, biological, social, and environmental factors (Muglia and Katz 2010; Conde-Agudelo et al. 2006; DeFranco et al. 2014) contributing to perinatal health.

8.2 Preterm Birth and Gestational Age Estimation

An infant born prior to the 37th week of pregnancy is considered preterm – though the earlier an infant is delivered, the greater the likelihood the infant will suffer complications. Preterm birth is the predominant contributor to infant mortality in the United States (Macdorman and Mathews 2008). At 11.5 %, the 2012 U.S. preterm birth rate remains above of the March of Dimes goal for 2020 of 9.6 % and far exceeds the March of Dimes aspirational goal for 2030 of 5.5 % (Hamilton et al. 2013; McCabe et al. 2014).

Gestational age is commonly used for categorization of neonatal cohorts and is a critical risk adjuster in calculating infant mortality and other neonatal outcomes. While most morbidity and mortality affects infants who are born extremely preterm (<28 weeks gestation) or very preterm (28 to <32 weeks gestation), even babies classified as late preterm infants (34–36 weeks gestation) are at elevated risk for neonatal morbidity and mortality and incur greater cost than their full term counterparts (McIntire and Leveno 2008; Kramer et al. 2000; Wang et al. 2004). Preterm infants are also at an increased risk to experience sudden infant death syndrome or other sleep related causes of death following their initial hospitalization (Task Force on Sudden Infant Death and Moon 2011; Malloy 2013). Gestational age is determined by the lapse of time following the onset of the mother's last menstrual period (LMP) prior to becoming pregnant, with the approximate normal gestational length lasting 40 weeks (see Fig. 8.2). Although precisely defined, there are many challenges to obtaining accurate and consistent measures of gestational age.

Several methods are available to estimate gestational age, yet the most commonly used is dating based upon the mothers LMP. Although defined by LMP onset, menstrual-based gestational age dating is error prone due to irregularities in women's cycle length and errors in accurate recall of related dates (Ananth 2007). Obstetricians are particularly interested in obtaining accurate gestational

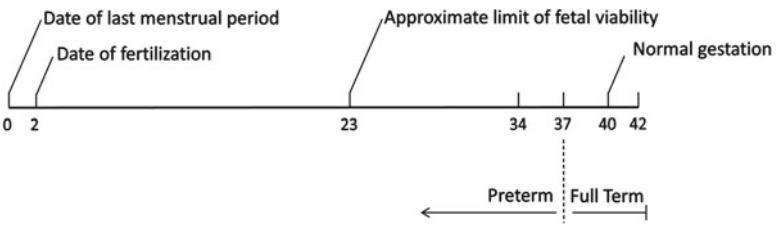


Fig. 8.2 Stages of pregnancy in weeks of gestation

age estimates to provide care appropriate to the level of fetal development. Obstetric estimates of gestational age are often based upon early ultrasound examinations. Within the first trimester, fetal growth is remarkably consistent, allowing for accurate estimation of gestational age based on ultrasound appearance. As pregnancy progresses beyond the first trimester, growth is less consistent, making gestational age assignment based upon infant appearance on ultrasound less reliable. When maternally reported LMP is dramatically different from ultrasound-based estimates of gestational age, the obstetrician must determine which estimate most likely represents the "true" gestational age of the infant. The obstetric estimate is typically recorded in the mother's medical record.

The Ballard or Dubowitz examinations can be performed on a newborn during the first day of life to estimate gestational age based on appearance, skin texture, motor function and reflexes. Although physical examinations have been validated for the entire newborn population including extremely preterm infants, the examinations are not routinely performed on full term infants (Ballard et al. 1991). Additionally, variation of 1–2 weeks of true gestational age is common, therefore physical examinations are more often used to confirm, rather than reassign, obstetric estimates of gestational age. Estimates derived from physical examinations will typically be recorded in the infant's medical record as they are obtained during postnatal care. Variations in timing and method may result in several discrepant estimates of an infant's gestational age being recorded in different charts or in different hospital units. To further complicate matters, United States vital statistics data also contain another representation of gestational age generated using an algorithm that mediates differences between LMP and obstetrician-based estimates using birth weight measures (Martin et al. 2002). Because the choice of variable used to represent gestational age may considerably impact calculated preterm birth rates, investigators and policy makers should aim for consistency in which representation is utilized (Hall et al. 2014b; Martin et al. 2015).

As an example of the importance of consistency in variable selection, a recent evaluation analyzed the impact of gestational age variable type on calculations of preterm birth at the population level (Hall et al. 2014b). The retrospective analysis reviewed all three methods of gestational age estimation available in Ohio vital birth statistics. Live birth records from 2006 to 2009 were compared and preterm birth rates were calculated using each gestational age variable. For each of 608,530 births, gestational age estimates based on LMP were compared to clinically-based obstetric estimates. When gestational age estimates did not perfectly agree, differences in the resultant classification of preterm birth status were evaluated with respect to the third, reconciliatory combined gestational age estimate. Substantial agreement was found in preterm classification among gestational age estimates (kappa: 0.748; 95 % Confidence Interval: 0.745–0.750); LMP-based gestational age estimates did not perfectly agree with obstetric estimates in approximately 40 % of records. Disagreement in gestational age led to disagreement in preterm birth classification in 5.3 % of total cases resulting in a 1.8 percentage point difference in preterm birth rate calculations (11.0 % using obstetric and 12.8 % using combined estimates). The analysis demonstrated a lack of agreement between gestational age

estimate variables that translated to a meaningful difference in classification of preterm birth rates and underscored the need for consistent use of gestational age measures in conducting population-level analyses of preterm birth or infant mortality.

Some studies use infant birth weight as a proxy for gestational age with ranges of birth weights assigned to various cohorts. Birth weights are nearly universally obtained and much less prone to error than gestational ages within a hospital setting; however, while birth weight is also an important data point to capture, there are limitations to using birth weight to represent fetal development. Infants who are either smaller or larger than appropriate for their gestational age may be at increased risk for poor outcomes; however, in the absence of gestational age information the birth weight loses some predictive power. Furthermore, gestational age is a measure not merely of size, but of the level of infant maturity and physiological development. Thus, gestational age becomes a useful guide in the appropriate prescription of medications and administration of therapies independent of birth weight.

8.3 The NICU and the Neonatal Electronic Health Record

Among preterm infants, immature lungs may lead to an array of respiratory complications often requiring mechanical ventilation or other respiratory support within the NICU setting. Premature infants are also more susceptible to infection, including pneumonia and sepsis, as well as tissue death of the bowels known as Necrotizing Enterocolitis (NEC). Immature digestive systems among preterm infants necessitate nutritional supplementation through intravenous feeding called Parenteral Nutrition (PN). Other conditions affecting both term as well as preterm infants, which are frequently managed within the NICU, include congenital anomalies requiring immediate surgical management, hyperbilirubinemia or jaundice, and neonatal abstinence syndrome or neonate withdrawal following intrauterine exposure to drugs or medications. Although very low birth weight and preterm infants make up the majority of admissions to NICUs, recent trends have demonstrated an increased risk for NICU admission for infants of all weight ranges (Harrison and Goodman 2015).

It is insufficient to merely install an Electronic Health Record (EHR) system designed for an adult or even general pediatric population into the NICU setting. In particular, the NICU EHR must support a higher level of granularity and precision of measurements than is traditionally available for more mature populations. Whereas in the adult ICU, an approximate patient age in years may suffice, neonatologists concern themselves with infant age in days and weeks of life. Similarly, an EHR capturing only the admission weight or periodically updated patient weight may lack the precision necessary to accommodate rapid infant growth. Particularly when medication dosages and nutritional orders are based upon the infant's weight, it is imperative that the EHR accurately represent the neonate's ever fluctuating weight. Further, because 10-fold weight differences comparing the smallest and the

largest infants admitted to NICU are common (i.e. 500 g vs 5 kg), accurate weights and medication dosing are essential to maintaining patient safety (Emmerson and Roberts 2013; Li et al. 2015). Additional considerations should be made in the neonatal EHR to support unique requirements for inpatient monitoring, newborn screens and neonatal specific scores, integration of growth charts, as well as integration with the maternal EHR (Dufendach and Lehmann 2015).

8.4 Perinatal Data Sources

Because perinatal health is influenced by a complex array of factors, researchers are interested in capturing a broad spectrum of relevant data. These data range from patient-specific measures to representations of worldwide health and include measures at numerous levels of granularity in between. Individual patient data may be captured by a number of sources, including care providers, billers, insurance providers, and quality improvement teams. These data include maternal and infant diagnoses, procedures, laboratory results, demographic and other characteristics, medication orders, treatment courses, intrauterine and postnatal environmental exposures, growth measures, outcomes, biospecimens, and genomic analyses. Although the categories of interest are not unique to the perinatal population, the sub-populations of newborns that attract the greatest research focus (including preterm infants) are defined by diagnoses and treatments not prevalent in other pediatric care domains.

Researchers are also interested in aggregated as well as area-level measures (see Fig. 8.3). Unit-level measures aggregate patient-level data at either the treatment unit/facility level or at the neighborhood/community level. For example, NICU investigators may be interested in studying patient costs, outcomes, or disease incidence within their hospital. Likewise, a health department may be interested in disease incidence or perinatal healthcare costs within a single community unit. Aggregations of data at the regional, national, or global level may provide important insights into assessing regional health quality and costs, identifying hotspots of disease activity or risk factors, or in guiding policy and lawmakers. Various levels of data aggregation provide the opportunity to combine health data with other regional datasets, including vital statistics, census, environmental, or other registries and databases that may support the investigation of population health hypotheses. In particular, community level perinatal studies may be strengthened by associating hospital acquired data with community datasets collected during follow-up care or by home visiting programs.

Quality of data and the purpose for which data are collected are important considerations when selecting which elements to include in analyses. For example, ICD-10 (International Classification of Diseases, Tenth Revision) codes are generated from discharge summaries and other clinical notes by non-clinical personnel

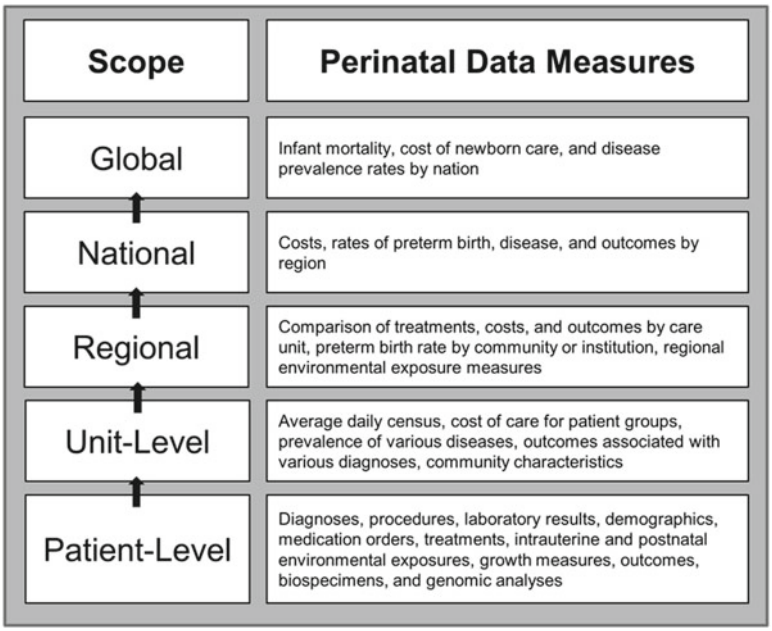


Fig. 8.3 Examples of perinatal data at various levels of aggregation, from patient-specific to global in scope

primarily for billing and other administrative purposes, not to support clinical care. Despite numerous analyses demonstrating the deficiencies in completeness, accuracy, granularity, precision, and content of ICD codes, they continue to be commonly used in research because they are discrete and easily acquired (Iezzoni et al. 1992; Hsia et al. 1988).

More specific to perinatal research are the limitations of birth certificate or vital statistics data. In the United States, birth certificates are the only population based, national data source of measures such as gestational age (Wier et al. 2007). Therefore, because the birth certificate data set captures more than 99 % of births annually, it remains an important resource in studying the health of maternal and newborn populations in spite of data quality concerns, which arise from unreliable source data and variation in data collection (Schoendorf and Branum 2006). The birth certificate data set contains large numbers of incomplete records particularly for records associated with high-risk women and vulnerable infants (Gould et al. 2002). While some birth certificate fields such as maternal risk factors and pregnancy complications are often under reported, other measures such as pregnancy history, outcomes, and demographics have been found to be reliable data sources (Vinikoor et al. 2010).

8.5 Perinatal Informatics Applications

Many of the breakthrough advances in neonatology occurred prior to the integration of computerization and medicine. Although a specialized subfield of neonatal or perinatal informatics has been slow to emerge, the hope by many is that the adoption of Information Technology (IT) systems, will "save babies" (Miller and Tucker 2011). There remains great potential for neonatologists and other perinatal care providers to build upon core concepts of Biomedical Informatics to improve perinatal care. (Drummond 2009) The American Medical Informatics Association defines Biomedical Informatics as, "the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving and decision making, motivated by efforts to improve human health" (Shortliffe 2010). Thus, neonatal or perinatal informatics likewise concerns itself with those same efforts related to the domain of infant health spanning from the molecular to the population level. Consequently, informatics applications to support improved morbidity and mortality among newborn populations can be focused within the newborn care setting itself or may take a broader, population-level, lifespan approach inclusive of women during preconception or pregnancy phases.

8.5.1 Computerization of Neonatal Care

The last few years have produced numerous neonatal endeavors building on core informatics concepts. Tools have been integrated into the NICU EHR to improve safety through better coordinated hand-off, computerized physician order entry, decision support for medication prescription, and detection of medication administration errors (Palma et al. 2011a, b; Li et al. 2014, 2015; Hum et al. 2014). Other efforts to improve the EHR have focused on human factors, seeking to improve the presentation of EHR data to better aid the physician in the decision-making process (Brown et al. 2011; Ellsworth et al. 2014). Using metadata produced by the EHR, others have sought to model and improve patient care teams (Gray et al. 2011). Analytical models have been developed to predict mortality among preterm infants (Medlock et al. 2011), as well as to predict demand for NICU resources (Khazaei et al. 2015). At the population health level, global infant trends and projections in mortality rates have been a focus of study (Wang et al. 2014). While tools are being integrated into hospital-based EHRs, other efforts have focused on web-based dissemination of decision support to standardize care processes and to improve the continuity of newborn care (Longhurst et al. 2009; Thornton et al. 2007). Signal analysis has been applied to interpreting newborn heart rate variability with applications including determining the severity of Hypoxic Ischemic Encephalopathy as well as to better monitor and understand the effects of ground and air transport on infants (Gholinezhadasnefestani et al. 2015; Karlsson et al. 2012). Additionally, the analyses of heart rate characteristics and EEG have aided in the prediction of neurodevelopmental outcomes among some groups of high acuity newborns (Addison

et al. 2009; Spitzmiller et al. 2007). Along with improving care quality and outcomes, it is hoped that technological advances will support cost savings.

Computerized NICU Tool Implementation Research projects at the Cincinnati Children's Hospital Medical Center (CCHMC) NICU aid in illustrating the implementation process of automated decision support tools within the inpatient setting. In Chap. 9 (Clinical Decision Support and Alerting Mechanisms) the authors describe a system being developed within the NICU EHR to support medication safety through real-time alerting of medication administration errors. The NICU setting was chosen in part because of the broad range of patient weights in the neonatal population which amplifies the severity of potential drug dosing errors (Li et al. 2015). While the system has great potential for improving neonatal outcomes, the framework is also adaptable to non-NICU settings through the development of unit specific algorithms.

Another system being developed to aid in the management of neonatal abstinence syndrome (NAS) is focused more specifically on improving neonatal outcomes. NAS following an infant's in-utero exposure to opioids has increased dramatically in recent years, affecting 5.8 per 1000 newborns nationally in 2012 (Patrick et al. 2012, 2015). As many as 60 % of infants with intrauterine opioid exposures experience withdrawal signs including hyperirritability, tremors, excessive crying, and seizures necessitating prolonged neonatal hospitalization and gradual weaning off opioids (Seligman et al. 2010). Clinical factors including type and extent of in-utero exposures have been identified as predictors for infant withdrawal severity, but genetic factors also play a key role in the presentation and severity of NAS (Kaltenbach et al. 2012; Wachman et al. 2013, 2014, 2015).

Machine learning algorithms are currently being developed to support personalized treatment for NAS based upon individual clinical and genetic factors [see Fig. 8.4]. In Cincinnati, mothers are universally tested for opioids at the time of delivery (Wexelblatt et al. 2015). As a consequence, newborns who have experi-

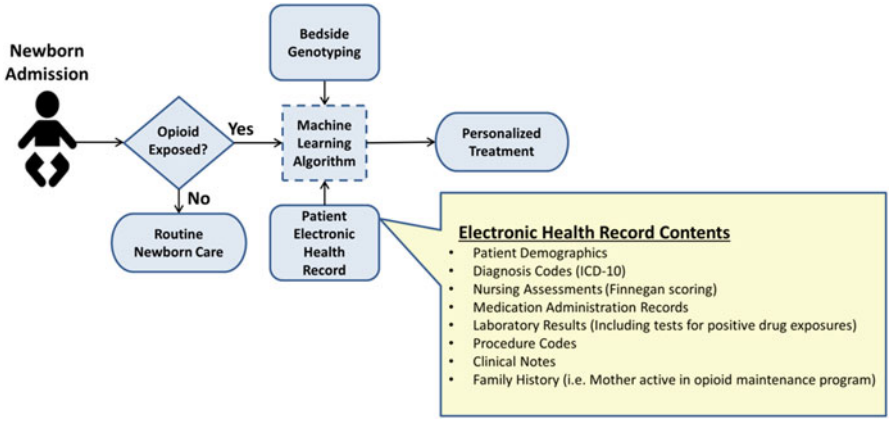


Fig. 8.4 Machine learning to enable personalized treatment of neonatal abstinence syndrome

enced intrauterine opioid exposure are rapidly identified and exposure is documented within the newborn EHR. Infants without opioid exposure receive routine newborn care; however, documentation of intrauterine opioid exposure will trigger the execution of an algorithm to support personalized treatment. Bedside genotyping will be performed for opioid exposed infants. Detection of genes associated with pharmacologic requirements or withdrawal severity will inform the machine learning algorithm, Chap. 17, (Genomics-based Stratification Strategies for Personalized Pain and Adverse Drug Effect Management in Pediatrics). Combined with clinical and demographic factors identified in the EHR (including newborn and maternal demographics, diagnosis codes, nursing assessments and Finnegan scores, medication administration records representing pharmacologic treatment course, laboratory results including positive tests for opioids and other substances including benzodiazepines, cocaine, methamphetamines, and marijuana, procedure codes, and family history including maternal self-reported drug exposures or reported activity in an opioid maintenance program) the algorithm will predict the need for pharmacologic treatment with a first-line weaning agent as well as secondary outcomes such as expected length of treatment and need for adjuvant drug therapy. Further, future refinements will enable the algorithm to recommend tailored pharmacologic treatment for NAS by identifying the most effective treatments (i.e. buprenorphine vs. methadone) associated with various individual clinical, genomic, and exposure profiles. As the major driver of cost to treat NAS is length of hospitalization (estimated at over $3000 daily), the implications of more timely and efficient treatment for reducing costs of NAS management could be substantial.

8.5.2 Informatics Support for Perinatal and Population Health

A number of barriers prevent the seamless integration of maternal and infant medical records needed for optimal clinical management for the mother-infant dyad. Although pertinent data exist at the time of birth, the newborn infant is essentially issued a blank medical record with a brand new medical record number (MRN). Maternal and fetal care including informative laboratory tests and ultrasound images provided up until birth, as well as documented pregnancy complications, are recorded in the mother's prenatal care record. Information related to the mother's prior pregnancy history and general medical history may be summarized in the prenatal record, but may also be detailed in separate maternal charts. Likewise, the paternal medical history is stored in the father's various medical records. A number of barriers prevent the seamless aggregation of data into the newborn's medical record including the transition between care institutions, transition between provider types, concerns for privacy, the absence of an established medical home, and technical limitations (See Chap. 1, Electronic Health in Pediatrics and Research and Chap. 2, Protecting Privacy in the Child's Electronic Health Record).

While information pertaining to the newborn delivery is captured at the birth hospital, prenatal data is typically collected in a clinic setting. Although prenatal

data may be obtained to guide delivery care, it is often in a non-computerizable (i.e. faxed, photocopied, or scanned) or summarized state, which does not facilitate integration with the hospital EHR. Transfers of a newborn to a higher acuity care center may further complicate data integration, particularly when the birth facility and transfer facility belong to different care networks or do not share an EHR system resulting in incongruent infant medical record numbers. Even the use of the same base EHR system by two facilities does not guarantee seamless exchange of patient data, even from a technological perspective. Additionally, large quantities of data may be missing and unobtainable in medical records belonging to infants of mothers who received scant or no prenatal care.

During pregnancy, care is predominantly delivered by an obstetrician or midwife. Although the health of both fetus and mother are of concern to the care provider, data capture is associated with the mother's MRN. Following delivery, newborn care delivered by neonatologists and pediatricians is recorded in the infant's own medical record represented by an MRN assigned to the infant. The transition of care providers accompanying the transition from fetal to infant stage results in a separation of data, although some modern EHRs are built to facilitate the manual migration of relevant data from the maternal to infant chart during the course of care.

As described in Chap. 2 on protecting privacy, concerns for maintaining data confidentiality must be satisfied when integrating newborn data with parental health information, including data pertaining to the newborn's prenatal care. While confidentiality remains a consideration during the course of administering clinical care, considerations for privacy are particularly stringent during the course of conducting research.

A final challenge to integration of perinatal data is the absence of a newborn's established medical home, which complicates the acquisition of follow-up data related to ultimate outcomes. Although the birth hospital or NICU care providers may inquire as to the intended primary care provider for the soon to be discharged infant, that patient provider relationship has often not been established. Furthermore, data integration once again is challenged by transitions in care providers and institutions.

Even if all other barriers are mitigated, linking newborn records remains a particular challenge, as many of the identifiers traditionally used by record matching algorithms are absent or unreliable in the newborn medical record. The newborn does not have a unique social security number or any other unique identifier that can be used in linking, nor in many cases does the infant have a name. Often the name listed on the newborn's medical record is represented by the same last name as the infant's mother and "baby boy" or "baby girl", if a first name has not yet been chosen. The infant name may be subsequently modified by providing a first name or by adjusting the last name, which further complicates linking. Probabilistic and deterministic matching of newborn records as well as newborn-to-mother records at the individual level are still possible using characteristics such as gender, date of birth / date of delivery, birth / delivery hospital, insurance type, and address elements such as street number and zip code (Hall et al. 2014c). Nevertheless, in cases of multiple

gestation births, detailed characteristics such as infant birth weight may be necessary to break ties between potentially matched records.

With origins in classical epidemiology, spatial analysis is another opportunity for studying and improving perinatal health at a population level (Rushton and Lolonis 1996). Traditionally, spatial analysis techniques have been utilized to identify geographical disease outbreaks. Population-based interventions may be designed around specific disease profiles in areas exhibiting excess disease prevalence (Caley 2004; Richards et al. 1999). Geocoding of patient address information and utilization of geographic information systems software facilitates the linkage of individual records to proximal area-level, census-based or environmental measures (Hall et al. 2014a; DeFranco et al. 2016). Geographical analyses have been used to study patterns related to pregnancy outcomes including stillbirth, preterm birth, birth defects, and other suboptimal perinatal outcomes associated with environmental exposures or neighborhood specific socio-demographic factors (English et al. 2003; DeFranco et al. 2015; Hall et al. 2012; South et al. 2012; Goyal et al. 2014).

8.6 Perinatal Population Systems

An example of a successful implementation for supporting population health research is the Pregnancy to Early Life Longitudinal (PELL) data system. The public-private collaborative partnership between the Maternal and Child Health Department at the Boston University School of Public Health, the Massachusetts Department of Public Health, and the Centers for Disease Control and Prevention has enabled researchers in Massachusetts to investigate a broad range of topics related to pregnancy, premature birth, and infant mortality (Barfield et al. 2008). The PELL system is a longitudinally linked data set of mothers and their children that allows researchers to track mothers and children over time. The core of the data system is a deterministically and probabilistically linked public health data set consisting of vital statistics, health services utilization data, and program participation data (Shapiro-Mendoza et al. 2006). The data core can subsequently be linked to additional resources for novel analyses. In addition to monitoring preterm birth rates and associated neonatal morbidities, PELL has been linked to ancillary databases to evaluate eligibility and referrals of children with developmental delay to an early intervention program to enable analyses of outcomes following assisted reproductive interventions to evaluate maternal outcomes following elective Cesarean delivery, and for surveillance of autism (Clements et al. 2008; Kotelchuck et al. 2014; Manning et al. 2011; Kim et al. 2015; Declercq et al. 2007). PELL also offers a conceptual model for other states hoping to conduct a similar range of analyses, including the opportunity to perform analysis related to specific mothers over multiple subsequent pregnancies.

In Wisconsin, a different approach to supporting perinatal research has taken the form of PeriData.Net. The web-based database was developed in partnership between the Wisconsin Association for Perinatal Care and the University of

Wisconsin Milwaukee/Center for Urban Population Health. Over 200 hospitals utilize the system to house hospital-supplied data corresponding to the state birth certificate fields. In addition to enabling electronic submission of birth certificate data, the system allows institutions to benchmark patient outcomes against other participants and to supplement the standard set of data fields with custom fields intended to support specific initiatives (Costakos 2006). Additionally, the system provides for standardized data capture facilitating the potential coordination of perinatal focused efforts between participating partners.

Unlike the systems in Massachusetts and Wisconsin which use vital records as the backbone for the data systems, a regional perinatal data system being developed in Cincinnati will leverage clinical and billing records. CCHMC physicians provide nearly comprehensive clinical coverage for newborns throughout the greater Cincinnati region as they are contracted to direct newborn care in each the region's delivery hospitals. Although an overwhelming majority of the infants seen by CCHMC neonatologists and pediatricians are seen in delivery hospitals, these newborn encounters generate physician billing records managed by the CCHMC EHR system resulting in a regional population-based data set. As a consequence, the CCHMC EHR newborn billing records are able to serve as a backbone for data linkage. Furthermore, the newborn billing records enable linkage to maternal and infant EHR records generated by each delivery hospital's EHR system as well as to vital statistics (see Fig. 8.5). Cincinnati investigators are hopeful that this Maternal and Infant Data Hub will serve as a central data repository enabling integration of lifespan data from numerous ancillary sources. These additional data sets may be linked at the individual record level or at the geospatial level and include data sets such as:

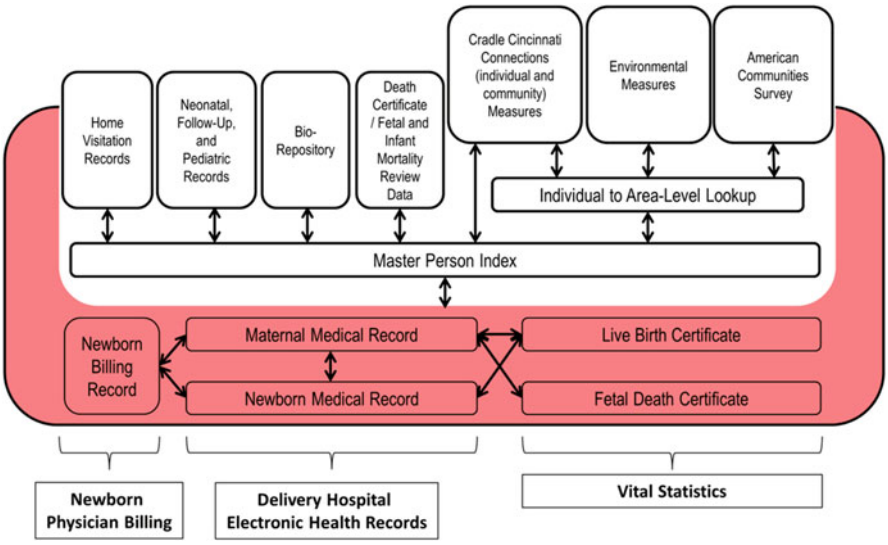


Fig. 8.5 Design of the Cincinnati maternal and infant data hub

- Community-based home visitation records
- Follow-up pediatric, urgent care, and emergency department visit data
- Hospital biorepository records
- Additional vital statistics, death records, and data from Fetal and Infant Mortality Reviews
- Environmental data including measures of airborne particulate matter (United States Environmental Protection Agency 2016)
- Census data (United States Census Bureau 2016)
- Data from other local perinatal focused programs such as Cradle Cincinnati Connections (Cradle Cincinnati 2014)

Investigators are hopeful that the integration of clinical EHR data will provide the Cincinnati system with a novel opportunity for studying perinatal health at the population level by enabling precise phenotyping and the comparison of clinical treatments and outcomes supporting more rigorous evaluations of public health interventions. Data governance remains a critical consideration in developing any population-based, shared data systems as institutional agreements must satisfy privacy, compliance, and data collaboration requirements.

8.7 Neonatal Research Resources

Many of the most clinically significant and costly conditions to treat in the neonatal population are relatively rare. For example, approximately 2.0 % of babies in the United States are very preterm, born at less than 32 weeks gestation. Necrotizing enterocolitis (NEC) affects about 10 % of the infants in the very preterm category. A single clinical site may not manage enough cases of NEC to power a statistically significant analysis of care interventions. For this purpose, multi-center research networks have been established to conduct clinical trials and observational studies in neonatal medicine.

The Vermont Oxford Network (VON) founded in 1986 and National Institute of Child Health and Human Development (NICHD) Neonatal Research Network founded in 1988 are two of the longest running quality improvement and research networks. VON maintains a database of information regarding high-risk newborn care and outcomes contributed by its more than 900 participating NICUs worldwide. The NICHD network also maintains a registry referred to as the generic database (GDB), which collects more detailed data than the VON database. In addition, the NICHD network coordinates research protocols for the 17 participating centers around drug therapies and therapies to manage neonatal conditions including sepsis, intraventricular brain hemorrhage, and chronic lung disease. Both networks focus primarily on very preterm infants with birth weights less than 1000–1500 g, as well as select other infants with specific conditions of high interest to researchers. These network databases enable the monitoring of disease trends and changes in neonatal survival and outcomes, while providing data for hypothesis generation and sample size calculation.

Another national scale database including data from over 4000 hospitals is the Kids Inpatient Database (KID). KID was established by the Healthcare Cost and Utilization Project (HCUP) sponsored by the Agency for Healthcare Research and Quality (AHRQ) to include a broad range of data collected during children's inpatient hospital encounters. KID includes between two and three million hospital discharges for children 20 years old and younger, a significant proportion of whom were newborns. The data including diagnoses, procedures, patient demographics, charges, length of stay, and hospital characteristics have been used to for analysis of national health care utilization, charges, quality, and outcomes (HCUP Kids' Inpatient Database (KID) 2016).

Although the NICHD and VON network collect similar measures from similar populations, the data dictionaries used by each database are not compatible. For example, both networks request information about retinopathy of prematurity (ROP). Vermont Oxford asks if a retinal exam was done and if so, requests the worst ROP stage documented on a scale of 0–5. Subsequently, yes or no responses are requested to inquiries of whether or not the ROP was treated with an anti-WEGF drug or surgery. The GDB also asks if ROP was diagnosed; however, rather than asking the ROP stage, the GDB requests to know if ROP reached stage 3 or worse. In addition to requesting information about treatment with anti-WEGF drugs, information about the specific intervention therapies of retinal ablation, scleral buckle, or vitrectomy is collected. For another example, both networks request data about necrotizing enterocolitis. The Vermont Oxford network asks for a yes or no response to the question of whether necrotizing enterocolitis occurred and if NEC was present, if surgery was performed. The GDB data requests NEC status as a single question including the options of "Absent/Suspect," "Proved, no surgery," or "Proved, surgery." In cases where definitions of the two networks are approximate, even in the absence of exact one-to-one mappings between data elements, it may be possible to translate data from one format to the other; however, incongruent granularity may result in information loss.

Although clinical terminologies such as SNOMED CT were designed to enable data standardization and exchange, existing terminologies do not adequately represent the pediatric domain, particularly concepts unique to neonatology. Insufficient established standard terminology has lead individual research networks to establish their own data definitions to achieve their research objectives. As a consequence the ability to pool or share data between research networks is severely limited. Furthermore, leveraging complimentary datasets such as the outcomes focused Children's Hospitals Neonatal Database (CHND), which captures therapies and outcomes of infants referred to tertiary centers, is also frustrated.

Development of the Neonatal Research Networks Terminology (NRNT) was motivated by the goal of enabling the exchange of perinatal data among research networks (Padula 2012; Kahn et al. 2014). Modeled after the structure of SNOMED CT and developed by subject matter experts, NRNT contains mappings to clinical findings, observations, and procedures used in current neonatal research networks. Serving as a hub, NRNT can facilitate meta-analyses as well as studies spanning

multiple research networks. By providing the structure of a central hierarchy of neonatal concepts, NRNT enables the translation of data from one network to another, as well as the aggregation of data for higher powered studies.

References

Addison K, et al. Heart rate characteristics and neurodevelopmental outcome in very low birth weight infants. J Perinatol. 2009;29(11):750–6.

Allen MC, et al. The limit of viability – neonatal outcome of infants born at 22 to 25 weeks' gestation. N Engl J Med. 1993;329(22):1597–601.

Ananth CV. Menstrual versus clinical estimate of gestational age dating in the United States: temporal trends and variability in indices of perinatal outcomes. Paediatr Perinat Epidemiol. 2007;21 Suppl 2:22–30.

Baines MA. Excessive infant-mortality; how can it be stayed? A paper contributed to the National Social Science Assn., London meeting. To which is added a short paper, Reprinted from the Lancet [1861] on Infant-Alimentation; or, Artificial Feeding, as a Substitute for Breast-Milk, Considered in Its Physical and Social Aspects. London: Churchill; 1862.

Ballard JL, et al. New ballard score, expanded to include extremely premature infants. J Pediatr. 1991;119(3):417–23.

Barfield WD, et al. Using linked data to assess patterns of Early Intervention (Ei) referral among very low birth weight infants. Matern Child Health J. 2008;12(1):24–33.

Blaxter M. The health of the children : a review of research on the place of health in cycles of disadvantage. Studies in deprivation and disadvantage. London: Heinemann Educational; 1981.

Brown P, et al. Variations in faculty assessment of Nicu flowsheet data: implications for electronic data display. Int J Med Inform. 2011;80(7):529–32.

Caley LM. Using geographic information systems to design population-based interventions. Public Health Nurs. 2004;21(6):547–54.

Clements KM, et al. Maternal socio-economic and race/ethnic characteristics associated with early intervention participation. Matern Child Health J. 2008;12(6):708–17.

Conde-Agudelo A, et al. Birth spacing and risk of adverse perinatal outcomes: a meta-analysis. JAMA. 2006;295(15):1809–23.

Costakos DT. Of lobsters, electronic medical records, and neonatal total parenteral nutrition. Pediatrics. 2006;117(2):e328–32.

Cradle Cincinnati. Serving moms in Cincinnati's west side. 2014. Accessed 2 Mar 2016. Available from: http://www.cradlecincinnati.org/connections-2/.

Declercq E, et al. Maternal outcomes associated with planned primary cesarean births compared with planned vaginal births. Obstet Gynecol. 2007;109(3):669–77.

DeFranco E, et al. Influence of interpregnancy interval on birth timing. BJOG. 2014;121(13):1633–40.

DeFranco E, et al. Air pollution and stillbirth risk: exposure to airborne particulate matter during pregnancy is associated with fetal death. PLoS One. 2015;10(3):e0120594.

DeFranco E, et al. Exposure to airborne particulate matter during pregnancy is associated with preterm birth: a population-based cohort study. Environ Health. 2016;15(1):6.

Drummond WH. Neonatal informatics–dream of a paperless Nicu: Part One: The emergence of neonatal informatics. NeoReviews. 2009;10:e480–e87.

Dufendach KR, Lehmann CU. Topics in neonatal informatics: essential functionalities of the neonatal electronic health record. NeoReviews. 2015;16(12):e668–e73.

Ellsworth MA, et al. Clinical data needs in the neonatal intensive care unit electronic medical record. BMC Med Inform Decis Mak. 2014;14:92.

Emmerson AJ, Roberts SA. Rounding of birth weights in a neonatal intensive care unit over 20 years: an analysis of a large cohort study. BMJ Open. 2013;3(12):e003650.
English PB, et al. Changes in the spatial pattern of low birth weight in a Southern California county: the role of individual and neighborhood level factors. Soc Sci Med. 2003;56(10):2073–88.
Gholinezhadasnefestani S, et al. Assessment of quality of Ecg for accurate estimation of heart rate variability in newborns. Engineering in Medicine and Biology Society (EMBC), 2015 37th annual international conference of the IEEE. IEEE, 2015.
Gould JB, et al. Incomplete birth certificates: a risk marker for infant mortality. Am J Public Health. 2002;92(1):79–81.
Goyal NK, et al. Association of maternal and community factors with enrollment in home visiting among at-risk, first-time mothers. Am J Public Health. 2014;104 Suppl 1:S144–51.
Gray JE, et al. Using digital crumbs from an electronic health record to identify, study and improve health care teams. AMIA Annu Symp Proc. 2011;2011:491–500.
Hall ES, et al. Spatial analysis in support of community health intervention strategies. AMIA Annu Symp Proc. 2012;2012:311–20.
Hall ES, et al. Integrating public data sets for analysis of maternal airborne environmental exposures and stillbirth. AMIA Annu Symp Proc. 2014a;2014:599–605.
Hall ES, et al. Evaluation of gestational age estimate method on the calculation of preterm birth rates. Matern Child Health J. 2014b;18(3):755–62.
Hall ES, et al. Development of a linked perinatal data resource from state administrative and community-based program data. Matern Child Health J. 2014c;18(1):316–25.
Hamilton BE, et al. Births: preliminary data for 2012. Natl Vital Stat Rep. 2013;62(3):1–20.
Harrison W, Goodman D. Epidemiologic trends in neonatal intensive care, 2007–2012. JAMA Pediatr. 2015;169(9):855–62.
HCUP Kids' Inpatient Database (KID). Healthcare cost and utilization project (Hcup). 2016. Accessed 2 Mar 2016. Available from: www.hcup-us.ahrq.gov/kidoverview.jsp.
Hsia DC, et al. Accuracy of diagnostic coding for medicare patients under the prospective-payment system. N Engl J Med. 1988;318(6):352–5.
Hum RS, et al. Developing clinical decision support within a commercial electronic health record system to improve antimicrobial prescribing in the neonatal icu. Appl Clin Inform. 2014;5(2):368–87.
Iezzoni LI, et al. Comorbidities, complications, and coding bias. Does the number of diagnosis codes matter in predicting in-hospital mortality? JAMA. 1992;267(16):2197–203.
Kahn MG, et al. Building a common pediatric research terminology for accelerating child health research. Pediatrics. 2014;133(3):516–25.
Kaltenbach K, et al. Predicting treatment for neonatal abstinence syndrome in infants born to women maintained on opioid agonist medication. Addiction. 2012;107 Suppl 1:45–52.
Karlsson BM, et al. Sound and vibration: effects on infants' heart rate and heart rate variability during neonatal transport. Acta Paediatr. 2012;101(2):148–54.
Khazaei H, et al. Health informatics for neonatal intensive care units: an analytical modeling perspective. Translational engineering in health and medicine. IEEE J. 2015;3:1–9.
Kim SY, et al. Prevalence of adverse pregnancy outcomes, by maternal diabetes status at first and second deliveries, Massachusetts, 1998–2007. Prev Chronic Dis. 2015;12:E218.
Kotelchuck M, et al. The mosart database: linking the Sart Cors clinical database to the population-based Massachusetts Pell reproductive public health data system. Matern Child Health J. 2014;18(9):2167–78.
Kramer MS, et al. The contribution of mild and moderate preterm birth to infant mortality. Fetal and infant health study group of the Canadian perinatal surveillance system. JAMA. 2000;284(7):843–9.
Li Q, et al. Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care. J Am Med Inform Assoc. 2014;21(5):776–84.

Li Q, et al. Automated detection of medication administration errors in neonatal intensive care. J Biomed Inform. 2015;57:124–133.
Longhurst C, et al. Development of a web-based decision support tool to increase use of neonatal hyperbilirubinemia guidelines. Jt Comm J Qual Patient Saf. 2009;35(5):256–62.
Macdorman MF, Mathews TJ. Recent trends in infant mortality in the United States. NCHS Data Brief. 2008;9:1–8.
Malloy MH. Prematurity and sudden infant death syndrome: United States 2005–2007. J Perinatol. 2013;33(6):470–5.
Manning SE, et al. Early diagnoses of autism spectrum disorders in Massachusetts birth cohorts, 2001–2005. Pediatrics. 2011;127(6):1043–51.
Martin JA, et al. Births: final data for 2001. Natl Vital Stat Rep. 2002;51(2):1–102.
Martin JA, et al. Measuring gestational age in vital statistics data: transitioning to the obstetric estimate. Natl Vital Stat Rep. 2015;64(5):1–20.
McCabe ER, et al. Fighting for the next generation: Us prematurity in 2030. Pediatrics. 2014;134(6):1193–9.
McIntire DD, Leveno KJ. Neonatal mortality and morbidity rates in late preterm births compared with births at term. Obstet Gynecol. 2008;111(1):35–41.
Medlock S, et al. Prediction of mortality in very premature infants: a systematic review of prediction models. PLoS One. 2011;6(9):e23441.
Miller AR, Tucker CE. Can health care information technology save babies? J Polit Econ. 2011;119(2):289–324.
Muglia LJ, Katz M. The enigma of spontaneous preterm birth. N Engl J Med. 2010;362(6):529–35.
Padula M. Neonatal research network terminology harmonization: a formative research initiative of the national children's study. National children's study metadata repository workshop. 2012.
Palma JP, et al. Neonatal informatics: computerized physician order entry. Neoreviews. 2011a;12:393–96.
Palma JP, et al. Impact of electronic medical record integration of a handoff tool on sign-out in a newborn intensive care unit. J Perinatol. 2011b;31(5):311–7.
Patrick SW, et al. Neonatal abstinence syndrome and associated health care expenditures: United States, 2000–2009. JAMA J Am Med Assoc. 2012;307(18):1934–40.
Patrick SW, et al. Increasing incidence and geographic distribution of neonatal abstinence syndrome: United States 2009 to 2012. J Perinatol. 2015;35(8):650–5.
Reidpath DD, Allotey P. Infant mortality rate as an indicator of population health. J Epidemiol Community Health. 2003;57(5):344–6.
Richards TB, et al. Geographic information systems and public health: mapping the future. Public Health Rep. 1999;114(4):359–73.
Rushton G, Lolonis P. Exploratory spatial analysis of birth defect rates in an urban population. Stat Med. 1996;15(7–9):717–26.
Saigal S, Doyle LW. An overview of mortality and sequelae of preterm birth from infancy to adulthood. Lancet. 2008;371(9608):261–9.
Schoendorf KC, Branum AM. The use of United States vital statistics in perinatal and obstetric research. Am J Obstet Gynecol. 2006;194(4):911–5.
Seligman NS, et al. Relationship between maternal methadone dose at delivery and neonatal abstinence syndrome. J Pediatr. 2010;157(3):428–33. 33 e1.
Shapiro-Mendoza CK, et al. Risk factors for neonatal morbidity and mortality among "healthy," late preterm newborns. Semin Perinatol. 2006;30(2):54–60.
Shortliffe EH. Biomedical informatics in the education of physicians. JAMA. 2010;304(11):1227–8.
South AP, et al. Spatial analysis of preterm birth demonstrates opportunities for targeted intervention. Matern Child Health J. 2012;16(2):470–8.
Spitzmiller RE, et al. Amplitude-integrated eeg is useful in predicting neurodevelopmental outcome in full-term infants with hypoxic-ischemic encephalopathy: a meta-analysis. J Child Neurol. 2007;22(9):1069–78.

Task Force on Sudden Infant Death, Syndrome, R. Y. Moon. Sids and other sleep-related infant deaths: expansion of recommendations for a safe infant sleeping environment. Pediatrics. 2011;128(5):1030–9.
Thornton SN, et al. Neonatal Bilirubin management as an implementation example of interdisciplinary continuum of care tools. AMIA Annu Symp Proc. 2007:726–30.
United Nations, Department of Economic and Social Affairs, Population Division. World population prospects: the 2010 revision, Cd-Rom Edition. 2011.
United States Census Bureau. American Community Survey (Acs). 2016. Accessed 2 Mar 2016. Available from: https://www.census.gov/programs-surveys/acs/.
United States Environmental Protection Agency. Airdata. 2016. Accessed 2 Mar 2016. Available from: http://www.epa.gov/airdata/.
Vinikoor LC, et al. Reliability of variables on the North Carolina birth certificate: a comparison with directly queried values from a cohort study. Paediatr Perinat Epidemiol. 2010;24(1):102–12.
Wachman EM, et al. Association of oprm1 and comt single-nucleotide polymorphisms with hospital length of stay and treatment of neonatal abstinence syndrome. JAMA. 2013;309(17):1821–7.
Wachman EM, et al. Epigenetic variation in the Mu-Opioid receptor gene in infants with neonatal abstinence syndrome. J Pediatr. 2014;165(3):472–8.
Wachman EM, et al. Variations in opioid receptor genes in neonatal abstinence syndrome. Drug Alcohol Depend. 2015.
Wang ML, et al. Clinical outcomes of near-term infants. Pediatrics. 2004;114(2):372–6.
Wang H, et al. Global, regional, and national levels of neonatal, infant, and under-5 mortality during 1990–2013: a systematic analysis for the global burden of disease study 2013. Lancet. 2014;384(9947):957–79.
Wexelblatt SL, et al. Universal maternal drug testing in a high-prevalence region of prescription opiate abuse. J Pediatr. 2015;166(3):582–6.
Wier ML, et al. Gestational age estimation on United States livebirth certificates: a historical overview. Paediatr Perinat Epidemiol. 2007;21 Suppl 2:4–12.

Chapter 9
Clinical Decision Support and Alerting Mechanisms

Judith W. Dexheimer, Philip Hagedorn, Eric S. Kirkendall, Michal Kouril, Thomas Minich, Rahul Damania, Joshua Courter, and S. Andrew Spooner

Abstract More than 55 % of US hospitals have electronic health records (EHRs); frequently these contain computerized decision support (CDS) in the form of alerts. Alerts are a common form of CDS often implemented for medication ordering and decision support to improve patient care. EHRs implement rules supplied by third-party vendors to help guide the dosing process include weight-based dosing. Since many of these rules are conservative, they result in noisy alerting and are therefore

J.W. Dexheimer, Ph.D. (✉)
Departments of Pediatrics and Biomedical Informatics, Divisions of Emergency Medicine and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Ave, ML-2008, Cincinnati, OH 45229, USA
e-mail: judith.dexheimer@cchmc.org

P. Hagedorn, M.D.
Department of Pediatrics, Division of Hospital Medicine, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Ave, ML-9016, Cincinnati, OH 45229, USA

E.S. Kirkendall, M.D.
Departments of Pediatrics and Biomedical Informatics, Divisions of Hospital Medicine and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Avenue, MLC-3024, Cincinnati, OH 45229, USA

M. Kouril, Ph.D.
Departments of Pediatrics and Biomedical Informatics, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Avenue, ML-7024, Cincinnati, OH 45229-3039, USA

T. Minich, RPh • J. Courter, Pharm.D.
Division of Pharmacy, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, ML-15010, Cincinnati, OH 45229, USA

R. Damania, B.S.
Northeast Ohio Medical University, 4209 OH-44, Rootstown, OH 44272, USA

S.A. Spooner, M.D., M.S., FAAP
Departments of Pediatrics and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Avenue, MLC-9009, Cincinnati, OH 45229, USA
e-mail: andrew.spooner@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_9

overridden by users. Alert fatigue is commonly studied and reported by providers. EHR implementers customize these rules to reduce noise. Adverse drug events are a common occurrence, prevalent in both adult and pediatric populations. However, there are few automated ways to identify adverse drug events. Weight-based dosing guidance for medication orders has limited functionality if the patient's body weight is entered incorrectly. Despite safeguards intended to prevent weight data-entry errors, erroneous weights exist in patients' charts. These pose a safety threat to patients, especially inpatients, whose medication doses may be calculated from the last recorded weight. In this chapter we will give an overview of pediatric clinical decision support in EHRs, the different modes and forms of CDS, as well as diving into several related and specific areas of CDS – medication dosing alerts and the detection of weight data entry errors. We will review these common sources of error and frustration in the EHR, how errors can be identified, and changes implemented to mitigate the errors.

Keywords Clinical decision support • Drug dosing • Electronic health records • Medication alerts

9.1 Electronic Health Records

Basic Electronic Health Records (EHRs) are installed in more than 75 % of hospitals across the United States (Charles et al. 2013). They are a major focus for research in health informatics (Brender et al. 2000). The EHR is a repository of patient data in a digital format (ISO/TR 2005). It encompasses secure storage and information that is accessible by providers. The goal is to deliver coordinated, high-quality, and safe care to patients. EHRs are used in both inpatient and outpatient settings, including healthcare professionals and administrative staff (Häyrinen et al. 2008). Recent implementations involve patients through patient portals in a more interactive system (Ball 2003; DeLeo et al. 1993).

To extend the utility of the EHR, functional requirements should be addressed. In 2001, the American Academy of Pediatrics (AAP) published a statement outlining features necessary for an EHR system to support health care for pediatrics. The AAP defined pediatric specific areas, which were desirable in the EHR. The emphasis was placed on three categories: data representation, data processing, and system design (AAP 2001). Newer statements (Dufendach et al. 2015; Kim and Lehmann 2008; Lehmann et al. 2015) define functional areas that are critical to the care of infants, children, and adolescents and that their absence results in impeding the quality of pediatric care. This ranges from immunization management to growth tracking to medication dosing (Spooner 2007).

Adoption rates remain low in pediatric hospitals with only 17.9 % of pediatric hospitals reporting having a basic EHR system and a similar percentage exchanged health information electronically (Nakamura et al. 2010). Furthermore, computerized provider order entry (CPOE), and clinical decision support (CDS) both believed to

be of essence for improving the quality and safety of care have been adopted by pediatric institutions (Chaudhry et al. 2006). Given this introduction to EHRs and a glance at the concept of CPOE and CDS in the pediatric setting, we can turn our attention to alerts in the EHR and common points of error.

9.2 Clinical Decision Support

Computerized decision support systems are integrated with EHRs to guide treatment decisions and to aid the decision-making process at the point of care. They frequently are built upon evidence-based clinical guidelines that represent the expert consensus on the best ways to manage patients under specific circumstances and help decrease variation in clinical practice (Bakken et al. 2004). Such decision support systems have demonstrated positive effects on patient outcomes (Dexter et al. 2004). For this reason, providers, payors, federal agencies, healthcare institutions, and patient organizations support the development, implementation, and application of decision support based on clinical guidelines. However, many barriers exist that limit the implementation and integration of these into clinical practice. In 1976, Clem McDonald noted the "non-perfectability of man" to emphasize the importance of reminding clinicians about patient- or disease-specific tasks during care (McDonald 1976).

Clinical decision support systems provide reminders and alerts about many elements of patient care, such as identifying potential medication warnings including drug-drug interactions, dosing suggestions, alerting about an abnormal laboratory value, recommending a preventive care measure or other care suggestions. At the simplest level, decision support systems, e.g., medication warnings and alerts, provide reminders to encourage "good" care and help increase the standard of care. These decision support elements are virtually invisible to the end-user because they are integrated with the software design of an EHR.

Decision support can be provided to users through passive or active alerts. Passive alerts are reminders that appear in a non-interruptive fashion as part of general care such as a drug-dosing suggestion or displaying allergy information. Active alerts provide a prompt or pop-up to the user that must be acted upon, such as when a new medication interacts with existing medications with a high probability of an adverse event. Active alerts are frequently time-dependent and strive to fit into the clinical workflows to display the right information at the right time to the right person. A potential issue with clinical decision support system alerts of all types is alert fatigue (Ash et al. 2007). Care must be taken with the design and implementation of context-specific alerts to reduce their number to those essential to providing good patient care and avoiding adverse events.

Clinical decision support systems are based on EHR data and generally require a knowledge base that represents the domain information and an inference mechanism, such as a rule engine, that is able to combine and evaluate patient information with information contained in the knowledge base. An example is avoidance of

adverse events by prospectively searching for potential drug-drug interactions, when a new medication is ordered in the EHR. The decision support system combines a patient's existing medication list with the new medication and executes rules that evaluates whether an interaction representation exists for the new medication and any of the other medications.

Decision support can be built and integrated with the medical record using rule-based or more complex techniques such as applying artificial intelligence methods (Friedman and Frank 1983). Artificial intelligence techniques such as Neural Networks, Bayesian networks, Natural Language Processing (discussed in Chaps. 11 and 12), or Support Vector Machines can be applied in domains, where more complex and multi-dimensional problems exist, such as classifying patients or suggesting likely diagnoses or treatment decisions. Many decision support approaches have sound theoretical approaches but are challenging to integrate in real-time patient care settings. Aspects such as workflow integration, alert-fatigue, specificity of alerts, or human-computer interaction characteristics all influence clinicians' acceptance of clinical decision support implementation (Aronsky et al. 2001).

9.3 Drug Dosing in Pediatrics

Pediatric patients are particularly vulnerable to iatrogenic harm. Several studies have shown higher rates of medical errors and adverse drug events (ADEs) in children than have been detected in adult populations. The reasons for this are numerous and are largely based on factors relating to physiologic growth and development. Some of the specific factors cited as increasing prescribing complexity are shown in Table 9.1. Due to the propensity for errors in this particular population, accurate clinical decision support is especially paramount.

Medication-related safety events, commonly known as ADEs, are a subset of all adverse events, are one of the most common types of errors. Due to their commonality, they warrant a little further discussion. ADEs are often assigned to a phase of the clinical medication-patient process. There are 5 phases of the medication process (Aspden et al. 2007):

(a) Ordering/prescribing
(b) Transcribing and verifying
(c) Dispensing and delivering
(d) Administering
(e) Monitoring and reporting

Most of the errors in pediatrics are related to the first phase, when care providers are prescribing medications. Drug-dosing makes up the largest proportion of prescribing errors, due in large part because dosing of children is based on their weight (such as 10 mg per kilogram of drug X) as opposed to absolute dosing used for most doses of adult medications (100 mg of drug X). As such, most of the literature on clinical decision support (CDS) in EHRs revolves around dosing support. One systematic review of the literature on CDS with medication dosing support showed

Table 9.1 Factors that increase the complexity of prescribing medications to a pediatric population (Kaushal et al. 2004; Koren et al. 1986; McPhillips et al. 2005 Feb; Rieder et al. 1988; Schirm and Tobi 2003; Zandieh et al. 2008)

Factor	Notes
Weight-based dosing	*Children are usually dosed in mg drug/kilogram weight as opposed to "fixed" absolute dosing in adults*
Varying drug metabolism & physiology	*As children develop, their physiologic properties change, which affects the pharmacokinetics and pharmacodynamics of drug metabolism*
Increase off-label use	*50–75 % of medications are labeled as having insufficient info for pediatric use* (Schirm and Tobi 2003)
Accurate/changing weight in growing children	*Weight often changes rapidly in children, which affects weight-based dosing*
Conversion of pounds to kilos	*Most parents relate their children's weight in the English system of weights (lbs. and ounces). Many prescribers and pharmacists must then convert that weight to the metric system (kilograms and grams). This conversion can create calculation errors.*
Many formulations, preparations, concentrations	*Medications often come in many different formulations with differing strengths. Oral forms often have several concentrations to consider.*
Total daily dose divided into multiple doses	*Pediatric dosing recommendations are often stated in total daily dosing, divided by a frequency of administration. This introduces chances of calculation errors.*
Tenfold errors can occur easily	*Pharmacists are less likely to recognize due to being used to adult doses* (Koren et al. 1986; Rieder et al. 1988)
Providers must know pediatric and adult dosing	*Older and/or larger children may need to be prescribed an adult-like absolute dose if the calculated weight-based dose exceeds a maximum amount*

that 95.7 % of studies pertained to either the prescribing or the monitoring phase (McKibbon et al. 2012). Drug dosing in pediatrics is discussed in more detail in Sect. 1.4.

9.4 Challenges Associated with CDS Implementation

Medication alerting is the primary form of CDS implementation at the point of ordering where, as described above, the pediatric population presents unique challenges to ordering providers and error prevention. Medication alerts generally encompass rules that apply to existing dosing, drug-allergy, drug-drug interaction, duplicate therapy and pregnancy/lactation guidelines and measures. During initial efforts aimed at systematic error reduction using CDS, medication alerts showed significant promise (Bates et al. 2001; Kaushal et al. 2001; Kirkendall et al. 2014). Early evidence of CDS systems, including medication alerts, appeared to show some benefit with reduction in medication error rates, but these early studies were comprised of a handful of "homegrown," or self-developed, systems and not vendor-based solutions which quickly gained momentum as adoption of CPOE spread

(Kaushal et al. 2003). Yet, there were warning signs that such broad system changes would carry unintended consequences and that CDS and medication alerting would not prove as sweepingly beneficial as hoped. Among the many limitations that emerged was the conclusion that providers often paid little, if any, attention to medication alerts with physicians overriding between 49 and 96 % of drug safety alerts according to some studies. High override rates were attributed not only to users, but systems which created errant or inappropriate alerts that ultimately led to distrust and/or alert fatigue (Van Der Sijs et al. 2006). While the phenomena of excessive alerting and alert fatigue has been well described, the ideal solution to this problem remains elusive (Ash et al. 2007).

One factor which contributes to the problem of excessive alerting and alert fatigue is the fact that pediatric drug dosing is often derived empirically based on adult dosing parameters (Barbour et al. 2014). However, the physiologic differences between neonates and adults necessitate different dosing strategies to achieve successful therapeutic outcomes while minimizing adverse events (Ivanovska et al. 2014). While decision support rules in CPOE systems attempt to create a consistent framework to evaluate appropriateness of medication orders, the complex nature of the theoretical dosing model is often unavoidably diluted when applied to the CPOE system (Stultz and Nahata 2015).

Taking these difficulties into account, current research and recommendations revolve around improving the signal-to-noise ratio in medication alerts. Approaches include a reasoned legal approach to alerting whereby regulation and standardization are utilized to reduce alerts to only high-risk scenarios as opposed to broad implementation of vendor-based rule sets (Kesselheim et al. 2011). Other efforts are aimed at standardizing alert content, appearance, language and usability as well as improving the clinical significance and specificity of alerts (Ayvaz et al. 2015; Beeler et al. 2014; Middleton et al. 2013). The former approach relies on reducing the number of overall alerts to reduce fatigue, while the latter hopes to achieve user buy-in by demonstrating value to the ordering provider.

The consequences of these varied approaches needs to be investigated more fully. For instance, what are the safety implications of broad deactivation of alerting rules except for those deemed most serious? Other possible areas of research include user response to more useable and actionable alerts. Would improving presentation and content convince practitioners that medication alerts are worthwhile or is this wishful thinking?

9.5 Medication Alert Analytics

Once implemented, alerts or clinical decision support of any kind should be evaluated for effectiveness. This is especially important in the current health information technology environment, which is heavily focused on not only facilitating patient care, but doing so safety and effectively, and with greater user satisfaction. Every month more data on the epidemiology and magnitude of issues like alert fatigue are

published in the informatics literature; these studies often demonstrate poor value and highlight the challenges of implementing CDS well.

Medication alert analysis can range from fairly simple techniques to very robust methodologies. Many EHR suppliers and third party vendors supply some tools to help customers understand how effective their alerts are, but often these tools are fairly rudimentary and are comprised of simple descriptive reports about the orders and alerts that were processed by their system. Often times these tools lack the granularity and clinical semantics to allow the in-depth analysis needed to address suboptimal configurations. In some cases, these reports may actually be misleading. For instance, recipients of these reports must have working knowledge of how the computerized provider order entry (CPOE) system processes orders to interpret the descriptive report output. Some configurations of CPOE systems may discontinue and reorder a medication when a user simply modifies the dose in the interface; this would generate a new order and, depending on your perspective, inflate the number of orders placed in the system, if the report and underlying analytics did not make this clear. Other CPOE systems simply modify the original order and only report both the original order and the modification as a single order. A comparison of order lifecycles across two different systems may lead investigators to erroneous conclusions about the ordering and alerting capacities of the organizations involved (an organization with lower medication order numbers may in fact have more orders than the one that initially appears to have more).

There are many different approaches to evaluating medication ordering and alerting. Each approach may have its own unique metrics, but there are some common measures that we will outline here. The most obvious metrics are the metrics mentioned in the previous paragraph; basic descriptive statistics around medication orders and medication-related alerts. Counts of medication and alert orders can be viewed and analyzed along many dimensions as shown in Table 9.2.

Metrics relating to rates (orders per patient encounter, alerts per 100 medication orders) are often used to normalize order and alert metrics so that variations in patient encounters, census, etc. do not lead to misinterpretations.

Table 9.2 Different dimensions and examples of ways to analyze medication orders and alerts

Dimension	Examples
Department	Surgical Services, Emergency Services
Location	Inpatient Unit A3, Medical Office Building Floor 1
Specialty	Internal Medicine, Endocrinology
Clinical role	Physician, Nurse, Occupational Therapist
Individual provider	Kirkendall, Eric
Temporal/time-relational	January 2017, 1st Quarter 2017, Day shift
Type	**Medication orders:** Narcotics, Opioids, Morphine, Pharmaceutical class, Therapeutic class
	Medication alerts: dosing, drug-drug interactions, allergy contraindications

Medication alerts have some specific metrics that are important to discuss. These alerts are often the product of one or more orders being processed by a rules engine in the EHR's CPOE system. It is very common for the EHR to record some aspects of the user's response to the alerts such as acceptance of the alert, overriding of the alert, or some other resultant user action due to the presentation of the alert. This data can be used to construct new metrics. One such metric is the alert salience rate:

$$\text{Saliance rate} = \frac{\text{Count of alerts cancelled or modified when alert presented}}{\text{Total number of alerts presented to the user}}.$$

Salience rate, described in earlier work, is the degree to which a user reacted (in a corrective manner) to an overdose alert (Kirkendall et al. 2014, 2016). This metric allows one to quickly relate the behavior-altering properties of the group of alerts being reported on. This rate can be followed over time as adjustments are made to the alert rules or engine, to gauge the clinical effects and safety of those changes. The inverse of the salience rate can be referred to as the alert override rate, the rate at which the user did not make any changes to the order(s) in response to a warning.

There are many other order and order alert metrics that can be created to address the analytic needs of any given investigation. The important principles to follow are: (1) the investigator consider the clinical situation, they are assessing, (2) the sources of the data, the data quality and limitations, and how the data is generated are understood, and (3) the investigator applies best statistical analysis practices when deriving their metrics, using absolute and relative methods when it makes sense to. By following these guidelines the data will provide revealing insights into how effective, or ineffective, the medication ordering and alerting clinical decision support tools are.

9.6 EHR Data Models

An important consideration in evaluating alerts is collecting and storing the EHR data. EHRs contain vast amounts of information about patient visits including clinical data from the patient visits from demographics, provider notes, orders and procedures, and all decision support metrics. EHR data models are typically highly normalized, which is not often suitable for targeted analytics. As a result, and to limit the data scope, one might utilize a data mart focused on the data of interest (e.g., orders – order_id, medication, dose value, dose units). Whether using the native SQL querying tool, or buying or building a reporting engine to interact with the database, it is important to clarify the meaning of "data" in the data mart. Although each EHR's reporting data warehouse comes with a data dictionary, it's often the case that not all information is captured. Hence it is critical to validate the data. That by itself can be a daunting task, but a simple spot-checking is of a value (e.g., is the date always in local time vs. UTC). One example of this is how Daylight

Savings Time is handled. For example for Daylight Savings Time, for seemingly overlapping orders, they might be 1 h apart, but a gap of 2 h between orders is really only 1 h during clinical care. One must take into account that EHRs have evolved over time and it's possible that the meaning of a particular data point might have evolved as well.

Another instance of the need for clearly defining data meaning is counting alerts. The system typically shows multiple alerts per window, but records each type of alert separately. Simple counting of rows in the alerts table might yield inflated numbers as those multiple alerts were displayed in a single screen to the ordering provider. Furthermore, a dosing alert might have multiple components (single dose overdose and daily overdose). Therefore, careful consideration of what we are looking for and how it is reflected in the data is important.

Further validation of the data is to compare the built-in EHR reports with results obtained through the data mart reports. Another validation step is to use a test environment and run through a few workflows subsequently comparing the expected data – e.g., finding the weight source when weight is used during alert determination for weight-based dosing.

Tracking down the evolution of information in EHRs, and how they changed over time, may be challenging, for instance if audit information for some data points are not available. Therefore, retrospective studies aiming to determine the system state at a given point in time could require significant effort. To address the lack of audit trails one could version the database and, instead of relying on the EHR built-in audit capabilities, the database level auditing could be used instead. This has some downsides, if the true reason for data change is not recorded.

Performance considerations become an issue as the number of analyzed data points increases. Fortunately, this can be addressed with modern relational database engines that are capable of handling millions of rows. With the proliferation of sensor data and ever increasing granularity the field is increasing moving toward distributed frameworks and database engines such as Spark, Hadoop, Cassandra, etc.

In our case we set out to analyze the impact of alerts on dosing safety. We related Alerts, Orders, Alert Rules, Demographics, Medications as well as Medication Administration data – each in its own table for agile querying. We utilize SQL Developer (Oracle™) and Tableau for visualization. We created a multidisciplinary team of physicians, pharmacists, and informaticists to help ensure data clarity and quality. In the end perfection is impossible, but as long as one can quantify the imperfection and report on shortcomings considerable improvement in understanding, your EHR's data can be achieved.

9.7 Patient-Attribute Errors

There are some EHR-based errors in clinical care that arise not directly from a user action, but due to how patient attributes have been recorded of handled in the EHR. In pediatrics, where body weight is of paramount importance in medication dosing, errors in the recording or processing of weight can create these latent errors,

whose effects do not occur until an order for a medication or radiation (Goske 2013) is written that depends on weight logic.

9.7.1 Weight Data Entry Errors

EHRs used in children should have the capacity to calculate drug doses based on body weight (Johnson et al. 2013; Spooner 2007; Stultz et al. 2015). Despite the ubiquity of the practice of weight-based dosing, there is remarkable variation across reference sources on what the correct dosage should be for a majority of pediatric drugs (Kirkendall et al. 2014). There are no reliable published estimates on the incidence of weight data-entry errors (Carpenter and Gorman 2002; Galanter et al. 2013), but one investigation estimated that about one in 3000 values are incorrect (Spooner et al. 2015). In an enterprise involving hundreds of thousands of patient encounters per year, many involving toxic medications, weight errors can be a serious threat to patient safety.

Several failure modes exist for weight errors:

- Confusion of pounds and kilograms (producing an error either 2.2 times smaller or larger than the real weight)
- Decimal errors (omission or misplacement) or confusion of kg and grams
- Mistyped digits ("fat finger" errors) or transpositions
- Entry of other vital-sign or anthropometric numbers (height, head circumference) in the weight field
- Entry of one patient's weight into another's due to opening the wrong chart
- Inappropriate handling of tare weight for medical equipment (e.g., wheelchairs)

The best defense against weight errors is consideration of the patient's weight vs. the historical pattern over time. Usually this is depicted in a growth chart, in which historical weights are plotted against age- and gender-based normative lines that give the user an idea of how likely that measurement is for an average child (CDC 2012). Alternatively, one may use percentile numbers for the current measurement, but this fails to put the number in historical context. Of course, it is not the "average child" that presents the most difficult challenge to detection of weight data-entry errors; it is the child with unusual patterns of growth (obesity, slow growth due to disease, etc.) that make growth norms useless for the detection of errors. Even so, there are challenges to current EHR medication order-entry design to make the growth chart visible during the ordering process. Detection of weight-based errors may be a fruitful area for machine-learning designs where unusual growth patterns can be distinguished from erroneous data entry.

Dose rounding (Johnson et al. 2011) is another weight-related, complex computational task that can result in errors. Rounding—especially for drugs in liquid form—must be done in order to create a feasible drug quantity for home administration. For example, a 2.2 mL dose might represent an accurate dose, but 2.0 or 2.5 mL

might better represent the dosing volume that a delivery device (syringe) would feasibly deliver.

9.7.2 Blood Pressure

Despite the high prevalence of electronic health records, the ubiquity of numeric blood pressure data, clear norms for their interpretation, and a clear guideline for management (Carroll 2015; National High Blood Pressure Education Program Working Group on High Blood Pressure in and Adolescents 2004), child health providers do poorly at identifying children with elevated blood pressure (Beacher et al. 2015; Hansen et al. 2007; Kaelber and Pickett 2009). The data suggests that only a minority of truly hypertensive children are recognized and treated or even follow up when their blood pressure falls outside of accepted norms (Brady et al. 2015). Part of the problem with this basic decision support task is that blood pressure norms depends on age, height, and gender. This complexity provides enough of a computational challenge that some EHR systems lack the ability to display whether a given blood pressure should be considered normal or not (Kaelber and Pickett 2009). Even in the case where such computation is supported, there is no standard way to display the data that has been proven to make a difference. It is still largely up to human diligence for high blood pressure in children to be detectable among the noise of data offered by the modern EHR.

9.7.3 Organ Dysfunction Metrics

In addition to weight-based dosing, there is a potential for adjustments due to organ dysfunction. When calculating kidney function, estimation of the glomerular filtration rate (GFR) is involved. In pediatric patients, the calculated estimate, according to the bedside Schwartz equation is directly proportionate to the patient's height (Schwartz et al. 2009). As estimated GFR is often used to decide upon the need for dose or frequency adjustment in cases of renal impairment, errors in height documentation could theoretically affect these decisions.

9.7.4 Weight-Based Dosing as a Function of Age

In addition to the need for doses based upon weight, the dose per patient weight changes as patients age. There are two general factors, which affect the rate of drug metabolism as children grow. The first of these is organ maturation, which is a function that asymptotically approaches adult function, and is near 90 % of mature function by the age of 2 years for most hepatic and renal functions (Kearns et al. 2003).

In addition to this, the function of organs is not directly related to total body mass, but is approximated by the ratios of weights to the power of 0.75 (Holford et al. 2013). The overall result of these changes can be seen with the dosing of phenobarbital, which decreases from 6–8 mg/kg/day to 4–6 mg/kg/day to 1–3 mg/kg/day in children 1–5 years, 5–12 years, and in adolescents, respectively (Lexi-Comp 2011).

9.8 Involving End-Users

Considering the growing concern over alarm fatigue and questions about the value of CDS alerts, it is increasingly important that proponents of alert refinement work to demonstrate their clinical impact. In other words, do improved alerts lead to better patient outcomes? One avenue of linking mediation alerts to outcomes is an algorithm based detection system, which uses medication alerts to prompt targeted chart review to examine association between a medication order and known adverse drug events (ADEs). Such a method would allow researchers and clinical operations to tailor alerts to clinically significant parameters, further improving specificity by presenting the user with alerts based on real-world, rather than theoretical, dosing rules.

There is no standard approach to evaluate and address these issues and data analysis alone has yet to reduce inappropriate alerts. Direct user report of erroneous medication rules may provide an opportunity to reduce clinically-inappropriate alerts. Users were asked to provide feedback in the modification and assessment of medication alerts in a pediatric emergency department (ED). We modified the medication order alert system interface in the EHR to incorporate feedback from prescribers in regards to the appropriateness of the alerts that were presented during clinical care. ED prescribers provided feedback for any alert they deemed inappropriate by either selecting the newly created override reason. We evaluated the feedback and corresponding medication dosing rules weekly and made changes to medication alerts. We provided feedback to the ED providers about any resulting modifications.

From 12/2013 to 3/2014, we received a total of 15 notifications from prescribers, containing 8 dosing-related notifications (Dexheimer et al. 2015). Of these, we modified 6 medication alerts (75 %); this resulted in an 86 % drop in alerts no longer being displayed per month. Alert modifications were made to match the pharmacy formulary recommendations and to accurately reflect ED protocols. Feedback to providers on all changes was well-received. Configuration changes to alert interfaces can help elicit feedback about medication alerts from prescribers at the point of care, and can lead to a reduction in the number of alerts presented to prescribers. This may help reduce alert fatigue and increase user trust in the system.

9.9 Future of Clinical Decision Support

The biggest barrier to effective clinical decision support is the distracting nature of the sheer volume of data that faces practitioners, when they turn to the computer system for information on their patients. We have known that computerized alerts have only limited effectiveness for a long time (McDonald 1976), and that has not changed over the years (Carroll 2015). Clearly, the model of thrusting unwanted information into the view of a clinician concentrating on the complexities of care is not a fruitful path for clinical informatics research. The grand challenges for clinical decision support in general have been well summarized elsewhere (Sittig et al. 2008); all of these apply in pediatric environments. Some challenges of specific import to decision support in pediatric care include:

- Moving to more standardized dose ranges for pediatric medications, including those which are widely used despite lack of governmental drug-agency approval in pediatric age groups
- Automating the judgment of the appropriateness of anthropometric measurements
- Validation of useful prediction heuristics for clinical deterioration in hospitalized infants and children
- Creation of mature data-sharing networks for rare pediatric diseases so that sufficient quantities of standardized EHR data can be used across multiple centers to derive valid conclusions about treatments and outcomes.

Since the adherence to alerts is multifactorial, moving forward research and operational work should be focused on ways to improve alerting methods including involving end-users in the design of alerts, evaluating rules used for triggering pop-ups, and the visual display of information. Alerting methodologies will remain in place and evaluating their effectiveness and improving their use could help improve clinical care and patient outcomes.

References

AAP. American Academy of Pediatrics: task force on medical informatics. Special requirements for electronic medical record systems in pediatrics. Pediatrics. 2001;108(2):513–5.

Aronsky D, Chan KJ, Haug PJ. Evaluation of a computerized diagnostic decision support system for patients with pneumonia. J Am Med Inform Assoc. 2001;8(5):473–85.

Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. AMIA Annu Symp Proc. 2007:26–30.

Aspden P, Institute of Medicine (U.S.), and Committee on Identifying and Preventing Medication Errors. Preventing medication errors. Quality chasm series. Washington, DC: National Academies Press; 2007.

Ayvaz S, Horn J, Hassanzadeh O, Zhu Q, Stan J, Tatonetti NP, Vilar S, Brochhausen M, Samwald M, Rastegar-Mojarad M. Toward a complete dataset of drug–drug interaction information from publicly available sources. J Biomed Inform. 2015;55:206–17.

Bakken S, Cimino JJ, Hripcsak G. Promoting patient safety and enabling evidence-based practice through informatics. Med Care. 2004;42(2):II-49–56.
Ball MJ. Hospital information systems: perspectives on problems and prospects, 1979 and 2002. Int J Med Inf. 2003;69(2):83–9.
Barbour AM, Fossler MJ, Barrett J. Practical considerations for dose selection in pediatric patients to ensure target exposure requirements. AAPS J. 2014;16(4):749–55.
Bates DW, Cohen M, Leape LL, Overhage JM, Shabot MM, Sheridan T. Reducing the frequency of errors in medicine using information technology. J Am Med Inform Assoc. 2001;8(4):299–308.
Beacher DR, Chang SZ, Rosen JS, Lipkin GS, McCarville MM, Quadri-Sheriff M, Kwon S, Lane JC, Binns HJ, Ariza AJ. Recognition of elevated blood pressure in an outpatient pediatric tertiary care setting. J Pediatr. 2015;166(5):1233–9. e1231.
Beeler PE, Bates DW, Hug BL. Clinical decision support systems. Swiss Med Wkly. 2014;144:w14073.
Brady TM, Neu AM, Miller 3rd ER, Appel LJ, Siberry GK, Solomon BS. Real-time electronic medical record alerts increase high blood pressure recognition in children. Clin Pediatr (Phila). 2015;54(7):667–75.
Brender J, Nøhr C, McNair P. Research needs and priorities in health informatics. Int J Med Inf. 2000;58:257–89.
Carpenter JD, Gorman PN. Using medication list – problem list mismatches as markers of potential error. Proc Amia Symp. 2002:106–110.
Carroll AE. How health information technology is failing to achieve its full potential. JAMA Pediatr. 2015;169(3):201–2.
CDC. Growth Charts. 2012. Retrieved from http://www.cdc.gov/growthcharts/.
Charles D, King J, Patel V, Furukawa MF. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2012: Office of the National Coordinator for Health Information Technology. 2013.
Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, Morton SC, Shekelle PG. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med. 2006;144(10):742–52.
DeLeo J, Pucino F, Calis K, Crawford K, Dorworth T, Gallelli J. Patient-interactive computer system for obtaining medication histories. Am J Health Syst Pharm. 1993;50(11):2348–52.
Dexheimer J, Kirkendall E, Kouril M, Mahdi M, Minich T, Spoonder S. Reduction of medication-related alerts through ED alert feedback at the point of care. Poster presented at the AAP 2015 national conference San Diego, CA. 2015.
Dexter PR, Perkins SM, Maharry KS, Jones K, McDonald CJ. Inpatient computer-based standing orders vs physician reminders to increase influenza and pneumococcal vaccination rates: a randomized trial. JAMA. 2004;292(19):2366–71.
Dufendach KR, Eichenberger JA, McPheeters ML, Temple MW, Bhatia HL, Alrifai MW, Potter SA, Weinberg ST, Johnson KB, Lehmann CU. AHRQ comparative effectiveness technical briefs. Core functionality in pediatric electronic health records. Rockville: Agency for Healthcare Research and Quality (US); 2015.
Friedman R, Frank A. Use of conditional rule structure to automate clinical decision support: a comparison of artificial intelligence and deterministic programming techniques. Comput Biomed Res. 1983;16(4):378–94.
Galanter W, Falck S, Burns M, Laragh M, Lambert BL. Indication-based prescribing prevents wrong-patient medication errors in computerized provider order entry (CPOE). J Am Med Inform Assoc. 2013;20(3):477–81.
Goske MJ. Image gently: child-sizing radiation dose for children. JAMA Pediatr. 2013;167(11):1083.
Hansen ML, Gunn PW, Kaelber DC. Underdiagnosis of hypertension in children and adolescents. JAMA. 2007;298(8):874–9.
Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inf. 2008;77(5):291–304.

Holford N, Heo YA, Anderson B. A pharmacokinetic standard for babies and adults. J Pharm Sci. 2013;102(9):2941–52.
ISO/TR. ISO/TR 20514:2005(en) health informatics— electronic health record— definition, scope and context. Geneva: ISO/TR; 2005.
Ivanovska V, Rademaker CM, van Dijk L, Mantel-Teeuwisse AK. Pediatric drug formulations: a review of challenges and progress. Pediatrics. 2014;134(2):361–72.
Johnson KB, Lee CK, Spooner SA, Davison CL, Helmke JS, Weinberg ST. Automated dose-rounding recommendations for pediatric medications. Pediatrics. 2011;128(2):e422–8.
Johnson KB, Lehmann CU, Council on Clinical Information Technology of the American Academy of, P. Electronic prescribing in pediatrics: toward safer and more effective medication management. Pediatrics. 2013;131(4):e1350–6.
Kaelber D, Pickett F. Simple table to identify children and adolescents needing further evaluation of blood pressure. Pediatrics. 2009;123(6):e972–4.
Kaushal R, Barker KN, Bates DW. How can information technology improve patient safety and reduce medication errors in children's health care? Arch Pediatr Adolesc Med. 2001;155(9):1002–7.
Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch Intern Med. 2003;163(12):1409–16.
Kaushal R, Jaggi T, Walsh K, Fortescue EB, Bates DW. Pediatric medication errors: what do we know? What gaps remain? Ambul Pediatr. 2004;4(1):73–81.
Kearns GL, Abdel-Rahman SM, Alander SW, Blowey DL, Leeder JS, Kauffman RE. Developmental pharmacology – drug disposition, action, and therapy in infants and children. N Engl J Med. 2003;349(12):1157–67.
Kesselheim AS, Cresswell K, Phansalkar S, Bates DW, Sheikh A. Clinical decision support systems could be modified to reduce 'alert fatigue' while still minimizing the risk of litigation. Health Aff (Millwood). 2011;30(12):2310–7.
Kim GR, Lehmann CU. Pediatric aspects of inpatient health information technology systems. Pediatrics. 2008;122(6):e1287–96.
Kirkendall E, Kouril M, Minich T, Spooner S. Analysis of electronic medication orders with large overdoses: opportunities for mitigating dosing errors. Appl Clin Inform. 2014;5:25–45.
Kirkendall ES, Kouril M, Dexheimer JW, Courter JD, Hagedorn P, Szczesniak R, Li D, Damania R, Minich T, Spooner SA. Automated identification of antibiotic overdoses and adverse drug events via analysis of prescribing alerts and medication administration records [Submitted for publication]. 2016.
Koren G, Barzilay Z, Greenwald M. Tenfold errors in administration of drug doses: a neglected iatrogenic disease in pediatrics. Pediatrics. 1986;77(6):848–9.
Lehmann CU, Weinberg ST, Alexander GM, Beyer EL, Del Beccaro MA, Francis AB, Handler EG, Johnson TD, Kirkendall ES, Lighter DE. Pediatric aspects of inpatient health information technology systems. Pediatrics. 2015;135(3):e756–68.
Lexi-Comp. "*Lexi-Comp*": Lexi-Comp OnlineTM, Lexi-Comp, Lexi-Drugs OnlineTM, and Ohio Hudson. 2011.
McDonald CJ. Protocol-based computer reminders, the quality of care and the non-perfectability of man. N Engl J Med. 1976;295(24):1351–5.
McKibbon KA, Lokker C, Handler SM, Dolovich LR, Holbrook AM, O'Reilly D, Tamblyn R, Hemens BJ, Basu R, Troyan S. The effectiveness of integrated health information technologies across the phases of medication management: a systematic review of randomized controlled trials. J Am Med Inform Assoc. 2012;19(1):22–30.
McPhillips H, Stille C, Smith D, Pearson J, Stull J, Hecht J, Andrade S, Miller M, Davis R. Methodological challenges in describing medication dosing errors in children. In: Henriksen K, Battles JB, Marks ES, Lewin DI, editors. Advances in patient safety : from research to implementation, Concepts and methodology, vol. 2. Rockville: Agency for Healthcare Research and Quality; 2005. p. 213–23.

Middleton B, Bloomrosen M, Dente MA, Hashmat B, Koppel R, Overhage JM, Payne TH, Rosenbloom ST, Weaver C, Zhang J. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc. 2013;20(e1):e2–8.

Nakamura MM, Ferris TG, DesRoches CM, Jha AK. Electronic health record adoption by children's hospitals in the United States. Arch Pediatr Adolesc Med. 2010;164(12):1145–51.

National High Blood Pressure Education Program Working Group on High Blood Pressure in, C. and Adolescents. The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. Pediatrics. 2004;114(2 Suppl 4th Report):555–76.

Rieder M, Goldstein D, Zinman H, Koren G. Tenfold errors in drug dosage. CMAJ. 1988;139(1):12.

Schirm E, Tobi H. Risk factors for unlicensed and off-label drug use in children outside the hospital. Pediatrics. 2003;111(2):291–5.

Schwartz GJ, Munoz A, Schneider MF, Mak RH, Kaskel F, Warady BA, Furth SL. New equations to estimate GFR in children with CKD. J Am Soc Nephrol. 2009;20(3):629–37.

Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, Campbell E, Bates DW. Grand challenges in clinical decision support. J Biomed Inform. 2008;41:387–92.

Spooner SA. Special requirements of electronic health record systems in pediatrics. Pediatrics. 2007;119(3):631–7.

Spooner SA, Dexheimer JW, Kouril M, Courter J, Hagedorn P, Damania R, Mahdi CM, Minich T, Kirkendall ES. Weight data-entry errors affect weight-based ordering and can pose risk to pediatric patients. Paper presented at the American Academy of Pediatrics National conference and Exposition, Washington, DC; 2015.

Stultz JS, Nahata MC. Complexities of clinical decision support illustrated by pediatric dosing alerts. Ann Pharmacother. 2015;49(11):1261–4.

Stultz JS, Porter K, Nahata MC. Prescription order risk factors for pediatric dosing alerts. Int J Med Inform. 2015;84(2):134–40.

Van Der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc. 2006;13(2):138–47.

Zandieh SO, Goldmann DA, Keohane CA, Yoon C, Bates DW, Kaushal R. Risk factors in preventable adverse drug events in pediatric outpatients. J Pediatr. 2008;152(2):225–31.

Chapter 10
Informatics to Support Learning Networks and Distributed Research Networks

Keith Marsolo

Abstract Many research and improvement activities, especially those that involve rare, pediatric, or chronic conditions, require the ability to pool, access or query data from multiple institutions. Here we describe informatics architectures that support quality improvement and research networks, or learning networks, as well as those that can support large-scale distributed research networks. Even though the activities and motivations of these networks are very different, they still require many of the same considerations in order to perform meaningful analysis on data that have been collected in multiple settings. This includes the measurement and characterization of data quality, the use of standardized or common data models, and the tracking and management of patient privacy, among others. We describe the informatics architectures of several learning networks, and two distributed research networks, detailing the commonalities and differences between them.

Keywords Distributed research networks • Learning health system • Learning networks • Multi-center quality improvement and research registries

10.1 Introduction

There is a push to transform traditional systems of healthcare into Learning Health Systems (LHSs), where new knowledge is quickly translated into general clinical practice and clinical practice serves as the engine to generate new evidence and knowledge, thus more effectively coordinating efforts between clinical care, quality improvement and research (C. Friedman and Rigby 2012; Friedman et al. 2010, 2015; Greene et al. 2012; Grossman et al. 2011a, b; McGinnis 2010; Olsen et al. 2011). Fundamental to a LHS is the learning health cycle (similar to the plan-do-study-act cycle), where users of the system (clinicians, patients, etc.) identify a

K. Marsolo, Ph.D. (✉)
Departments of Pediatrics and Biomedical Informatics, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Avenue, MLC-7024, Cincinnati, OH 45229, USA
e-mail: keith.marsolo@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_10

problem and assemble, analyze, and interpret data in order to generate feedback (e.g., reports, decision support) that leads to improvement (Cleghorn and Headrick 1996).

Two types of networks that support aspects of the LHS are learning networks and distributed research networks (DRNs). Learning networks have evolved from quality improvement and research networks, with pediatric sub-specialty networks as a prime example. Most pediatric chronic conditions meet the NIH definition for a rare disease (National Institutes of Health 2016), and no single care center has a sufficient number of patients to produce generalizable knowledge, a barrier that, unless addressed by networks or multi-center studies, slows the pace of knowledge acquisition and outcomes improvement. The American Board of Pediatrics (ABP) has supported the establishment of sub-specialty improvement and research networks in all 13 pediatric sub-specialties. This effort extends the successes of pediatric clinical networks that use data for improvement and (increasingly) research. Examples include: (1) the *Vermont Oxford Network (VON)* (est. 1988) that is dedicated to improving the quality of neonatal ICU care and (2) the *Children' s Oncology Group (COG)* (est. 1998), which focuses on clinical trials of new therapies, as well as studies of how to improve the delivery of existing therapies for pediatric cancer (Ross et al. 1993). The dramatic improvement in cancer survival rates from <10 % in 1962 to over 80 % today is a reflection of the benefits of an infrastructure that can systematically standardize care and enroll patients in research studies (Hunger et al. 2012). CancerLinQ is an example of an adult-focused oncology network (Shah et al. 2016; Sledge et al. 2013). Vast improvements in remission rate have been seen in networks like ImproveCareNow, which is focused on improving the care and outcomes of children and adolescents with Inflammatory Bowel Disease (IBD) (Crandall et al. 2011, 2012; Forrest et al. 2014b). The common themes drawn from collaborative pediatric learning networks that have demonstrated marked improvement in the outcomes of children with chronic disease are: (a) an unrelenting commitment to collecting high quality data, (b) continuously evaluating and proving their value to clinicians making in-the-trenches decisions, and (c) the long-term engagement of the participants and their institutions to sustaining the network (Forrest et al. 2014b)

Sub-specialty research and improvement networks offer advantages that are foundational for research "laboratories" (Margolis et al. 2009). Creation of total population registries at each site provide large and diverse study samples. By standardizing practice, they reduce variations in outcomes that are caused by variations in the way care is delivered, thereby increasing statistical power. By linking research to care delivery and engaging clinicians directly, networks provide a forum for user-led comparative effectiveness research (CER) (Selby et al. 2015), a core attribute of the learning healthcare systems (Olsen et al. 2007, 2011). Despite the pressing need to scale these research networks, expansion is hampered by the lack of an informatics infrastructure capable of supporting needed expansions to include a large number of participating clinical sites, while reducing the costs of conducting research. Sub-specialty research networks typically collect data manually, using specially designed disease-specific forms. As a result, they do not capture data directly in

electronic health records or incorporate the data into federated data warehouses. Furthermore, they do not make the data available to other researchers, allowing the data to be used to address a wide variety of secondary questions. These manual methods are costly, typically paper based, ad hoc (i.e., may depend on grant funding), and slow the pace of discovery by creating silos of data that are difficult or impossible to integrate.

Distributed research networks are gradually replacing the traditional central data coordinating center model because of privacy concerns in sharing large amounts of data from large patient populations. A small, but growing, number of distributed research networks have been implemented across a spectrum of clinical conditions and practice settings (J. S. Brown et al. 2010; Buetow and Niederhuber 2009; Fleurence et al. 2014; Libby et al. 2010; Loeffler and Winter 2007; Maro et al. 2009; Pace et al. 2009a, b; The caBIG Strategic Planning Workspace 2007; Toh et al. 2011). Recent years have seen the establishment of DRNs that are national in scale, including the network-of-networks National Patient-Centered Clinical Research Network (PCORnet) (Fleurence et al. 2014), which consists of both Patient-Powered Research Networks (PCORnet PPRN Consortium et al. 2014) and Clinical Data Research Networks (Devoe et al. 2014; Forrest et al. 2014c; Kaushal et al. 2014; Khurshid et al. 2014; Mandl et al. 2014; McGlynn et al. 2014; Ohno-Machado et al. 2014; Rosenbloom et al. 2014; Waitman et al. 2014), the Health Care Systems Research Network (HCSRN; formerly the Health Maintenance Organization Research Network (Moulton 1999; Steiner et al. 2014; Vogt et al. 2004)) and the Accrual for Clinical Trials (ACT) Network (National Center for Advancing Translational Sciences 2015). The characteristics of a DRN is that data stays local to the data partner and/or the partner retains control; data have been transformed into a common representation (see Chap. 6); data have been characterized and deemed "fit" for their intended use; there is a mechanism to distribute queries and analyses to participating partners; and partners generally have autonomy in deciding whether to execute a given query and return the results (Brown et al. 2009). While they do not have the same focus on improvement and care management as many learning networks, they provide the opportunity to run large-scale, population-level analyses over large numbers of patients.

The types of networks described above are ones where the primary source of data is a healthcare or claims provider. There is an emerging group of networks, as illustrated by the PCORnet PPRNs, where the primary source of data is the patient (Chung et al. 2016; Randell et al. 2014). These networks have the potential to collect much more richer data in areas like patient-reported outcomes, and can take advantage of the motivation of highly engaged patients. Their architectures and approaches are fairly disparate, however, so a discussion on these networks will be considered out of scope for the time being.

10.2 General Network Activities

While learning networks and DRNs are very different in their scope and overall approach, a key component to both is network governance. This includes items like an authorship policy, agreements on data use, operating principles, and the use of the network. Examples of such governance activities are summarized in Table 10.1.

Activities within a learning network generally fall into three broad categories: data collection, support of care management and quality improvement, and the support of research. Examples of the informatics solutions that enable these activities is provided in Table 10.2 and described in further detail in Sect. 10.5.

The informatics activities of a DRN are more straightforward than with a learning network, but include: the definition or adoption of a common data model, the adoption of extract-transform-load (ETL) conventions or specifications, a process to characterize or assess data quality, and a process or technology to distribute queries and return results. These activities are described in further detail in Sect. 10.6.2.

Table 10.1 Example governance activities

Item	Description
Network participation agreement/ operating principles	Covers purpose of the network, scope, expectations, how data will be collected, shared and used, etc.
Research/authorship policy	Indicates how research proposals are vetted/approved within the network, and steps for receiving authorship credit in a given study
Data use/business associate agreements	Legal agreements that are necessary before data can be shared between institutions for research and non-research purposes (data use and business associate, respectively)
Policies and procedures (e.g., use of network)	Describe the general operation of the network

Table 10.2 Informatics modules to support activities within a learning network

Activity	Informatics modules
Data Collection	Direct data entry (web forms)
	Data transfer/data upload
	Integration of external sources (e.g., EHR, PROs, claims, etc.)
Support of care management and quality improvement	Computable definitions/derived data processing
	Automated generation of reports
	Advanced analytics
	Data quality assessment
Research	Computable definitions/derived data processing
	Generation of datasets for analysis
	Ability to identify patients for trials

10.3 Approach

Learning networks utilize a variety of informatics architectures. Many of them operate a centralized registry, though there are examples of networks that utilize a distributed architecture, such as DartNet (Pace et al. 2009a, b). One of the reasons for choosing a centralized model relates to cost. It is often more cost effective for a network to have a single coordinating center that receives data from all participating centers. The data that are collected are standardized, with consistent definitions. Analysts at the coordinating center can become familiar with this information and provide guidance to participants on best practices for capture and collection. Another important factor is that most learning networks are focused on the collection of "standard of care" data, which is a smaller subset than might be collected in a traditional research registry. In contrast, DRNs usually have a more expansive scope in regards to data, dealing with a broader range of domains. In addition, instead of collecting a targeted, standardized set of elements with agreed upon definitions, centers are asked to map all of their existing data to a common representation. There may not be a 1:1 mapping between the source and the target, which means centers are forced to make a judgment on how to construct the mapping. As a result, it can be beneficial to have the data stay at the originating center, where there are analysts with strong knowledge of how the data were generated.

Another important distinction between learning networks and DRNs that influences their approach is that while DRNs are focused almost solely on research, learning networks have a heavy emphasis on quality improvement and/or care management. Care teams often make clinical decisions using QI and care management tools, so there is a need to have more up-to-date information than is required for research. This latency factor is another reason why learning networks use a smaller, more targeted set of data domains.

10.4 Privacy

Privacy concerns also play a role in the architecture of these networks. Most DRNs attempt to operate on data from all, or almost all, patients of a given data partner. The creation of an integrated data repository is generally viewed as human subjects research, which would require an IRB protocol, as this integration requires the compilation of data containing PHI. Obtaining consent on all of these patients is impractical, so most data partners request a waiver of consent in their protocols. Once the repository has been created, it is possible to execute queries against that return aggregate counts or summary statistics without additional IRB approvals, though the queries must generally adhere to the operating principles of the network. If a DRN chooses to launch a prospective study, data partners will obtain consent to collect additional information from study participants.

Learning networks, on the other hand, limit themselves to patients with a specific disease or condition of interest. Some networks utilize a model where data on the entire patient population (including PHI) are utilized for QI, clinical care and non-human subjects research activities (Marsolo et al. 2015; Shah et al. 2016). Consent is then obtained in order to use data for human subjects research. In contrast with a DRN, IRBs tend to not be in favor of a waiver of consent for a learning network because the members of participating care teams typically have a direct relationship with the patients, and therefore have several opportunities to obtain consent. While obtaining consent requires more work on the part of the care teams, it can help increase patient engagement with the activities of the network, and can allow for additional types of research that are not possible under the DRN model, including the ability to link and share data containing PHI with external parties.

10.5 Architecture and Activities – Learning Networks

There are many different possible informatics architectures that a learning network can utilize. A strawman architecture, which is based on the registry utilized by the ImproveCareNow Network (Marsolo et al. 2015), is shown in Fig. 10.1. In this architecture, the modules that support data collection and data entry are shown on the left. The outputs, which are how users would interact with the data of the registry, are shown on the right. All of the analytical processing occurs in the center. In this example, a series of procedures are executed on the raw input data to compute the derived or calculated variables that are needed to support the quality improvement, care management and research activities of the registry. These data are stored in an analytical data warehouse with periodic refreshes to other analytical data structures such as a data cube. The warehouse and other structures are then used to feed the outputs of the system.

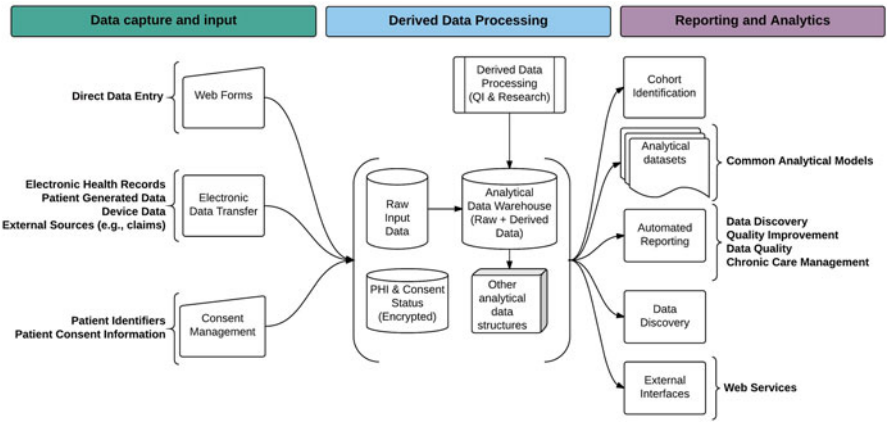


Fig. 10.1 Example architecture for a learning network

Inputs to the registry include data that are entered manually via web forms (e.g., visit information, laboratory results and medication orders, as well as population-based measures) and data that are transferred electronically using an automated or semi-automated process (e.g., data from electronic health records, patient-reported outcomes, claims and medication fulfillment information). It is also anticipated that the registry will be used to track information about patient consent status and the various identifiers used to manage patients and to link data that are obtained from multiple sources.

Outputs of the registry include automated reports that are used to support care management (pre-visit planning and population management), quality improvement and data quality. The registry should also support tools for data discovery, which would allow users to run ad hoc queries, slice-and-dice, and drill-down into various aspects of the data. Additional outputs include analytical datasets, such as those used to support comparative effectiveness research or common data models utilized by distributed research networks such as the PCORnet. Additional outputs include tools that allow users to quickly identify cohorts of interest and feeds to external applications, which could include patient-facing tools, mobile applications or even other registries. It is also expected that the registry will have interfaces to other tools such as a biorepository and clinical trial management system. The general requirements of each module or activity are described in further detail below.

10.5.1 Data Collection/Data Transfer

Data would be received by the system via either direct data entry or electronic transfer. Direct data entry would be handled by the Web Forms module, which would support the creation of network-specific web forms that allow for skip logic, conditional fields, edit checks, range values, etc. In a more advanced state, the Web Forms module would also include a Form Builder that allows general users to create web forms or modify existing forms to add/modify/delete individual fields without having to involve an application developer.

The registry would also support the upload of data from different sources. Examples sources include the EHR, patient reported outcomes (PROs), data on medication fulfillment, medical claims, etc. Where applicable, the registry would also support the ability to have uploaded data prepopulate the relevant fields of the web forms. While the Electronic Data Transfer module would define at least one standard input format for each data source/domain, to be most flexible for participants, it would also allow for some variation, for instance, to allow for EHR data from multiple vendors.

Modern EHRs support the creation of study-specific or condition-specific data collection forms. While most EHRs provide a number of different ways to capture data within the system, to be broadly useful for a network, EHR data collection form should the following general requirements. It should be possible to capture

responses as discrete data (i.e., not as part of a text string), clinicians should be able to pull the form responses into a progress note or referral letter, and it must be possible to extract the form responses from the system, either as a flat file or through another electronic interface, such as a web service or HL7 interface. If possible, the EHR vendor should be responsible for ensuring that the form questions and responses are mapped to standard terminologies where possible, though the network can also do those tasks. In an ideal state, the EHR form would be created so that the same form definition can be distributed and configured by any customer of that vendor, removing the need for centers to develop the forms locally.

In this model, each center generates an extract that is then sent to the network system on a periodic basis. This workflow is shown in Fig. 10.2. There are other alternatives to this approach, such as having an appliance be located behind the firewall at each institution, with a connection to that institution's EHR reporting database (or a database view into a subset of the overall database). In using an appliance, the data transfer process can be more straightforward, but there will be increased technical support costs on both the network and the institution, and it will likely require the use of additional legal agreements. While the use of other approaches to transfer data such as HL7 interfaces or EHR-based web services are appealing, the authors' experience is that the fastest and most cost effective way to obtain information from the EHR is through a flat file or similar extract.

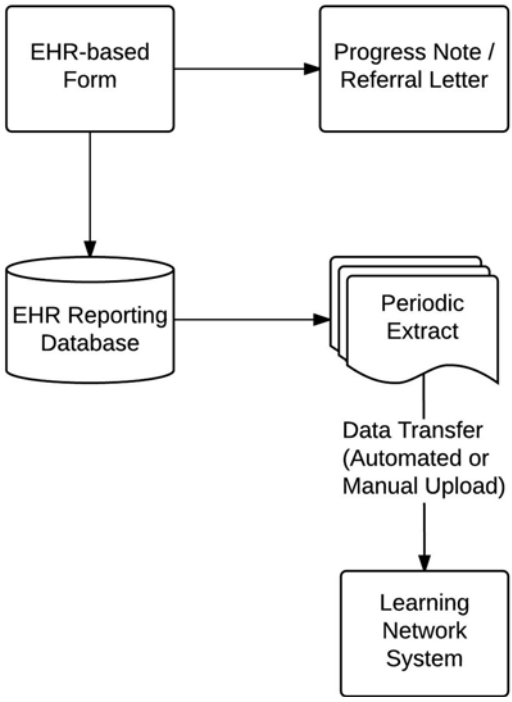


Fig. 10.2 Collecting data in the EHR and transferring it to the network system

In future years, there may be alternative approaches to the use of EHR-based data collection forms. This includes having patients provide more data, either through tablets or mobile devices in the waiting room, or by filling out forms that are sent to them through the EHR patient portal (see Sect. 10.5.3). Other potential options include workflows like those proposed by the Structured Data Capture Initiative (Office of the National Coordinator for Health Information Technology). In this scenario, a form is created and stored in an external form library. The EHR can be configured to download this form from the library, have it pre-populated with any relevant data (e.g., demographics) and then present it to the user for completion. Once the form is submitted, the data are sent to an external repository and (optionally) saved in the EHR. If widely adopted, the Structured Data Capture standard would remove the need to create vendor-specific EHR forms and allow the EHR to better support a number of ancillary activities, such as public health reporting, chronic disease registries and clinical trials. As it does rely on more complicated electronic interfaces to exchange data (e.g., web services), it is likely that there will be initial challenges in gaining access to the appropriate resources.

10.5.2 Device Data

In many chronic conditions, important outcome data are captured via medical device. In Type 1 Diabetes (T1D) for instance, most patient track blood glucose using pumps, meters, or continuous glucose monitors (CGMs). While it is usually possible to download data from these devices, each vendor tends to provide their own proprietary software/cable, requiring care centers to become familiar with many different interfaces and workflows. As a result, the process to pull data from a device ends up consuming a significant portion of the patient's clinic visit. Several 3rd party vendors are seeking to create universal uploaders that can talk to many different pumps and devices. The use of such an uploader provides several potential benefits, including having a single process for uploading data, common displays and metrics that can be used to interpret the data, and a single interface for sending the information to other data sources, such as a registry or EHR.

If it is important to have some form of the device data in the system in order to support the care management, improvement and research activities of the network, it is likely that it will also be important to have a version of the same data as part of the patient's medical record. Most commercial EHRs support the transfer of device or other patient-entered data into the EHR, if those data can be transferred using a standard HL7 interface. As a result, this presents two possible paths for the transfer of device data:

- Option 1: Device ->Uploader that supports HL7 ->EHR ->registry
- Option 2: Device ->Uploader that supports HL7 ->registry ->EHR

While working to integrate device data into the EHR may be appealing for an individual center, it is probably not a viable strategy for the network as a whole, for

two reasons. First, it takes a fairly significant commitment on the part of a local care center to support 3rd party integration (see previous section), and there are few economies of scale in terms of effort at the care center level. Second, there is some uncertainty as to the best way to interpret the device data when making clinical decisions. It is generally believed that summarized or synthesized interpretations of device data are more informative than the raw data itself, particularly when dealing with streaming devices like CGMs, but the optimal display is still unknown. If the device data were sent directly to the EHR, local EHR analysts would need to then create summary reports. With the current generation of EHRs, that effort would need to be repeated at every single care center. As a result, the most appropriate path of the device data in the short and medium-term may be option #2. Device data would be transferred to the registry first, where it would be summarized and visualized (depending on the sophistication of the uploader, this step could also occur outside of the registry), and then uploaded to the EHR, most likely as a scanned result. For care management and decision purposes, however, it would be optimal to have the raw device data and summary statistics eventually be included as part of the EHR, so networks should empower a small number of care centers to pilot direct integration with the EHR and use those findings to inform the rest of the network.

10.5.3 Data Integration – Patient-Generated Data (PGD)

The collection and transfer of patient-generation data presents many of the same challenges as that of device data. There is a lack of standardization in the types of data that may be collected – general surveys, validated patient-reported outcomes (e.g., PROMIS measures (Bevans et al. 2014; Forrest et al. 2014a, 2016; Gershon et al. 2010)), patient-specific outcome measures – and a lack of standardization on how those data are collected and stored (e.g., web-based surveys versus mobile applications). While PROMIS items can be represented as LOINC observations, allowing these data to be transferred to the EHR as standard HL7 messages, EHRs themselves have better support for directly capturing patient-generated data than device data. This includes the use of questionnaires within a patient portal, which may be web-based and/or support mobile devices, as well as through the use of kiosks or tablets within the clinic waiting room. As with device data, the same questions arise in how to interpret the information, whether to utilize the raw values or rely on a summarized version. As a result, the considerations and tradeoffs on how to handle patient-generated data within the system are largely the same.

10.5.4 Support for Chronic Care Management

Chronic disease registries like the one used by ImproveCareNow provide users with the ability to generate chronic care management reports (Marsolo et al. 2015). When used in conjunction with the rest of the interventions listed in the Chronic

Care Model (Bodenheimer et al. 2002; Coleman et al. 2009), these tools can help lead to improved patient outcomes. The chronic care management activities of the network would be supported by the system's Automated Reporting module. While this module could allow the generation of a number of different reports, it would be expected to include pre-visit planning and population management reports, as these are vital components of the chronic care model. Population management reports are used to help clinicians identify subpopulations. They provides aggregate information on each center's patient population according to metrics like demographics, medication usage, clinical status, risk assessment and patient-reported outcomes (if available). They also allow users to drill down into each metric and view longitudinal patient-level data on all patients that match the selected criteria. In an advanced state, centers would have the ability to create their own views or configurations within the report, which would allow them to specify the metrics they wished to include in their population-level graphs or the elements in the patient-level breakdown.

Pre-visit planning reports are used to help care teams plan upcoming clinic visits. The reports provide a snapshot of the patient's current status, their longitudinal history and help ensure patients are receiving proper medication dosing. They often include information on diagnosis and disease phenotype; selected information from past visits; patient-reported outcomes; a patient's current risk assessment; and considerations (recommendations) for medication dosing, lab ordering, and other actions based on the severity of disease.

Within the ImproveCareNow registry, the chronic care management reports can be generated on-demand and are refreshed on a daily basis. It is believed that these reports would be most effective to care teams if they were integrated into the EHR, which would allow them to immediately jump into a patient's chart and take action (e.g., schedule an appointment with a specialist, pend orders, etc.). It is possible to create care management reports within the EHR, but setup and configuration is time-consuming, requiring a significant amount of local effort. Since configuration is center-specific, there are few economies of scale. Emerging standards (Health eDecisions, Quality Reporting Document Architecture (Office of the National Coordinator for Health Information Technology; "Standards & Interoperability (S&I) Framework – Health eDecisions Homepage," 2013)) may eventually allow configurations to be shared, but at this point, it is infeasible to consider this as a scalable approach.

10.5.5 Quality Improvement/Data Quality

The Automated Reporting module should also include the ability to generate quality improvement and data quality reports. These reports would be presented as a series of run charts and control charts, dashboards, and report types like small multiples and spark lines. In an advanced state, these reports would be updated at least daily, and would be linked with the other chronic care management reports, for instance, allowing a center to pull up a cohort through their population management report

and view the process or outcome measures for that subset of patients. The advanced state would also include the ability for users (or superusers) to manage report metadata (e.g., centerlines, annotations, etc.). When viewing the data quality reports, it should also be possible to patient or visit level exception reports, so that users can quickly view and correct data entry errors.

10.5.6 Derived Data Processing/Advanced Analytics

The processes to compute derived data elements (e.g., process and outcome measures, derived variables that are used in other analytical activities, etc.) are crucial to supporting improvement and research activities, as well as in supporting care management decisions. In order to provide the most up-to-date information to end users, these processes should be automated and run in as close to real-time as possible. They should also allow variables to be computed at various time points (e.g., time of visit, as of the end of a month, as of a specific day). In an ideal state, the calculation process would be modular, allowing individual variables to be computed as the underlying data are changed, as opposed to have to run the entire process every time. To build trust in the system and enable transparency, the algorithms used to compute each measure or variable should be visible to the end users of the system.

The optimal data structure will depends on the needs of each network, but a single source of truth is generally preferred to having separate analytical datasets for each output or use case. A central dimensional data warehouse could be used for this purpose, periodically refreshing other analytical data structures, such as a multidimensional cube, which would provide functionality for ad hoc data discovery – e.g., the ability to drill-down or slice-and-dice variables of interest. More advanced analytical functions include subgroup analysis or the ability to identify groups of patients based on various certain similarity measures.

10.5.7 Cohort Identification/Patient Recruitment

One of the first steps in many research projects or clinical trials is the ability to quickly identify patients or cohorts of interest. There are a number of tools that can be used for such activities (see Chap. 6), but an important distinction is that the cohort identification tool should be able to query both the raw data, as well as those that are computed through the derived data process, as it may be necessary to identify patients using outcome or process measure values. In addition, while a self-service cohort identification tool may be used to satisfy simple queries, more complex inclusion and exclusion criteria will require that an analyst mediate the query. The tools should also support the ability to re-identify patients for potential contact, though those processes are also governed by the network's IRB protocol(s) and legal agreements.

10.5.8 Data Standards/Common Data Models/External Interfaces

To support large-scale comparative and observational research, the system should support the ability to output data in a variety of analytical formats or common data models (see Chap. 6). The support of additional outputs, such as web services, either custom-developed or adhering to standards like the Fast Healthcare Interoperability Resource (FHIR) (Health Level 7 2015; Raths 2014), would allow the system to interface with external applications. To support these activities, the various data domains in the registry (e.g., diagnoses, demographics, laboratory results, patient-reported outcomes) should be mapped to standard terminologies, wherever possible. For certain domains with a long history of local implementations, such as laboratory values, this standardization process may not be feasible for all possible results, and may require significant local effort on the part of each care center (Raebel et al. 2014). As government regulations such as Meaningful Use promote the adoption of standards and their use is increasingly tied to reimbursement (Blumenthal and Tavenner 2010), these challenges may become less of an issue. In the short-term, however, networks will need to choose how much effort they want to spend on data harmonization.

10.5.9 Management of Protected Health Information/Linkage to External Data Sources

Since most learning networks support chronic care management activities, there is a need to generate reports that contain PHI (Shah et al. 2016). The ability to collect, store and utilize this information will typically be governed by the network's IRB protocol(s) and legal agreements. To minimize risk, access to this information should be as limited as possible, and ideally, it would be encrypted, both in motion and at rest. In order to better support research and enable the collection of "complete" data on participants, the system should support processes that allow the use of PHI for secure linkage to external data sources with appropriate regulatory approval. These methods, which often create hash-based identifiers using the underlying PHI of the patient, are referred to as privacy-preserving record linkage techniques (Kho et al. 2015; Nasseh et al. 2014; Schmidlin et al. 2015), and can help augment existing registry data with information from other sources (e.g., claims and fulfillment data).

10.5.10 Management of Patient Consents/Electronic Consent (e-consent)

As learning networks often participate in multiple studies, having a way to centrally manage patient consent information can help determine which patients are participating in various studies. Stand-alone applications have been developed that allow

patients to record their consent decisions (Marsolo and Nix 2014), and integrating similar functionality into the system would be useful. At its most basic, the module to manage patient consent would store demographic information on participants, track their consent status, and send alerts to the study staff when a participant's consent expires. The ability to obtain consent electronically (e-consent) is also gaining traction, as witnessed by the quick adoption of Apple's Research Kit (Bot et al. 2016). Most e-consent tools have been targeted to adult participants. While pediatric e-consent tools exist, they involve more complicated workflows, due to the need to handle both consent and assent decisions.

10.5.11 Patient Access to Data

As illustrated by the establishment of organizations like the Patient-Centered Outcomes Research Institute (PCORI), recent years have seen a push to make the research process more patient-centered (Selby et al. 2012), which ranges from focusing on the outcomes and questions that matter most to patients, to including them at all steps of the research process, from design to dissemination. As a result, there is an increased desire for learning networks to provide the ability for patients to access their own data and results. Enabling such functionality presents a number of technical issues related to user management that have the potential to significantly raise the overall support costs of the technology platform. The staff at the care centers may also face an increased support burden, as patients will consider them to be the first line of technical support if they have problems accessing the system.

If a network wished to provide patients with access, it would need to decide whether patients will register directly with the system or whether patients will request access through their care centers. If it is the former, a process would need to be created in order to properly authenticate the identify of the patient and their relationship(s) to participating care centers, and if it is the latter, a process would be needed to handle patients moving between care centers. In addition, the network would need to decide whether patients are simply provided with access to their raw data, or whether some the systems will provide additional interpretation to give patients recommendations on potential courses of action. If such feedback is provided, the system may then be considered to be a diagnostic or medical device, which could potentially be regulated by government bodies like the FDA. A much simpler approach to providing patients with access to their data is simply for care centers to generate reports for each patient. The reports can be provided to patients during their clinic visit or be e-mailed to them prior to an appointment. If desired, care centers could also download a patient's data and provide it to them in a raw format. Such a process is less efficient, but likely to be more cost-effective, at least in the short-term.

10.5.12 Knowledge Management

Knowledge repositories can be used to share tools and process between network participants, increasing overall engagement and spurring the adoption of best practices. In order to be most effective, the content on these sharing platforms needs to properly tagged or cataloged so it can be indexed and searched. In their most advanced state, knowledge repositories can recommend appropriate tags based on the metadata of the material being uploaded. Otherwise, users must supply these tags as they upload content and/or a curator needs to follow-up and manually add this information after the fact. Since the goal of most knowledge repositories is to make it easy to share as much content as possible, the workflow for adding tags should not be too burdensome. The network can generate a pre-defined list of tags, which can periodically be refined to reflect new themes that may arise. If a network does not have the resources to add a curator to monitor the content on the knowledge repository, other staff members can be tasked with performing an audit of a small number of items to determine whether material is being appropriately cataloged. Training can be provided to users if it turns out that the wrong tags are used. It may also be possible to recruit highly motivated volunteer curators to perform this service for free, much as Wikipedia has relied on the efforts of a small number of editors to curate the site. For a network to fully take advantage of the material that is being shared, there should be links between the knowledge repository and the rest of the system. For instance, if a user is looking at a list of process failures on their quality improvement reports, the system should recommend tools or interventions from the knowledge repository that could help the user address such failures.

10.6 Architecture and Activities – Distributed Research Networks

The informatics architectures that support DRNs generally fall into two different categories: synchronous and asynchronous. In a synchronous network, queries are distributed and results returned in near-real time, without the need for partner-specific approvals. As a result, the potential throughput of a synchronous network is much higher than one that relies on an asynchronous architecture. The drawback is that queries on a synchronous network are typically less sophisticated than what is possible in an asynchronous one, often limited to simple counts and Boolean algebra. This is due to the fact that data partners do not have the ability to review results before they are sent back to the requestor, which means there is less trust in allowing sophisticated analyses to run without review. In addition, it may not be possible to run a sophisticated analysis in a matter of seconds, which is necessary when operating in a real-time fashion. Finally, the technical support costs are often higher for these architectures, as staff need to make sure that the network infrastructure is constantly running and available for query.

An example of a synchronous DRN architecture is the Shared Health Research Information NEtwork (SHRINE) (Weber et al. 2009), which is based on tools developed for the Shared Pathology Informatics Network (SPIN) (Drake et al. 2007; Namini et al. 2004). Examples of networks that leverage SHRINE are the ACT Network, UC Rex, SCILHS, and GPC (Mandl et al. 2014; National Center for Advancing Translational Sciences 2015; Waitman et al. 2014). While these networks may vary slightly in their topology, most implementations utilize a single Query Aggregator and a set of SHRINE adaptors, one for each repository in the network. Authorized researchers can query the network for the total number of patients at participating institutions who meet a given set of criteria—a combination of demographics, diagnoses, medications and laboratory tests. Because counts are aggregate, patient privacy is protected. The Query Aggregator consists of a user interface and a set of web services that broadcast queries to each SHRINE adaptor. A SHRINE adaptor translates each SHRINE query into the corresponding local terminology and queries the attached data repository. The local results are translated into the SHRINE terminology and sent back to the Query Aggregator, which compiles the results as a set of aggregated counts (some data partners elect to simply transform their data to conform to the shared ontology). While SHRINE adaptors already exist for the i2b2 research patient data warehouse (see Chap. 6, (Kohane et al. 2012; Murphy et al. 2010)), it is feasible to create adaptors to run on other data models, provided that the queries have been converted to run locally and a mapping has been created between the network terminology and the local codes. One consideration with SHRINE is that since queries are executed in real-time, the speed at which results are returned may be limited to the capacity of the slowest network member. While SHRINE has the ability to "time out" a partner after a certain period if there is no response, as the number of nodes on the network grows, there is a greater chance that a given partner is going to be having issues with connectivity or uptime. To date, the largest distributed SHRINE networks have less than a couple dozen members. There have been larger implementations (Marsolo et al. 2015; Natter et al. 2012), but these tend to run within a single data center.

Asynchronous architectures can potentially support more advanced analytics because there are not expectations that results will be returned in near real-time. An analysis can be distributed as code to be executed against a data partner's repository, with the results submitted back to the requestor after the data partner has a chance to review the results. While the time lag to return results is much greater (while it depends on the governance of the network, it is typically on the magnitude of several days), the results can be more sophisticated and complex. There will be manual effort on the part of the analyst, however, as they will need to download the code, execute it locally, and then return the results. An example of software the supports an asynchronous model is PopMedNet, which was created to facilitate comparative effectiveness research (PopMedNet 2011). While PopMedNet supports the concept of a "menu driven query" that can return aggregate counts using basic inclusion and exclusion criteria (e.g., demographics and diagnosis), its major feature is the ability to distribute and run executable analyses, which is illustrated in Fig. 10.3.

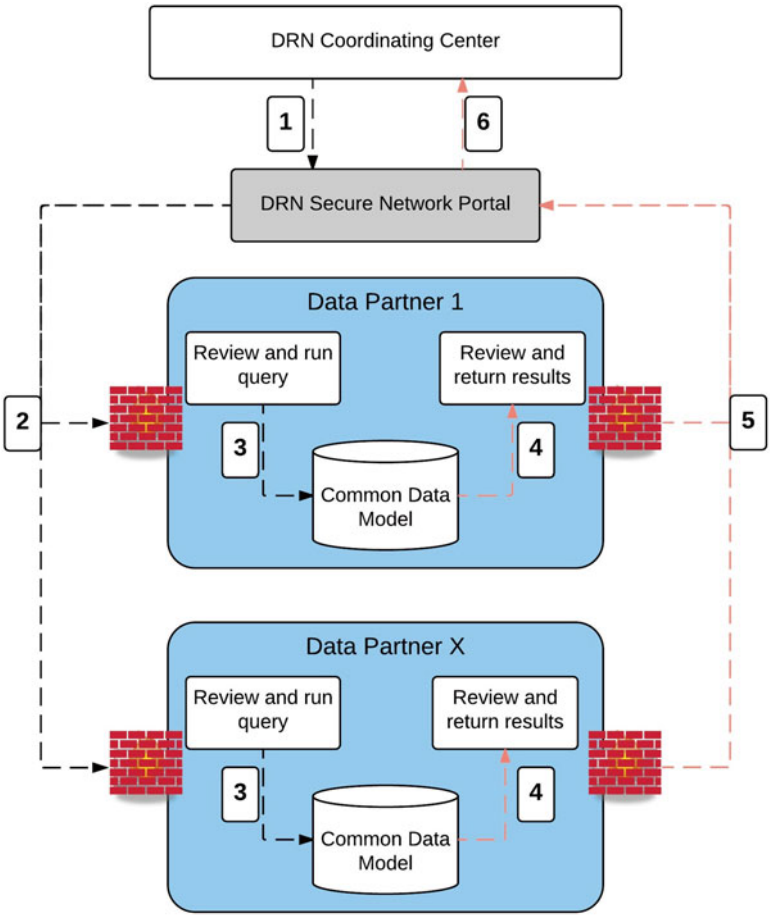


Fig. 10.3 Example DRN workflow using popmednet (**1**) User creates and submits query (executable code); (**2**) Data partners retrieve query from portal; (**3**) Partners review and run query against their local data; (**4**) Partners review results; (**5**) Partners return results via secure network; (**6**) Results are aggregated at the DRN portal

PopMedNet is used by the FDA Sentinel (formerly Mini-Sentinel) project, and is the platform that underlies PCORnet.

The decision on which architecture to adopt will ultimately depend on the purpose and aims of the network. If a network is limited to basic queries for cohort identification, a tool like SHRINE may be sufficient. If more advanced analysis is required, if a network involves dozens or hundreds of sites, or if there are concerns about infrastructure and maintenance costs, then PopMedNet may be a better fit.

As discussed in Sect. 10.2, there are a number of activities that must occur to successfully operate a DRN. The informatics activities can generally be categorized into three broad groups: develop and refine an extract-transform-load (ETL) con-

ventions document; adopt and execute data characterization routines or data quality assessments (DQAs); and iterate and expand the DRN common data model. Each of these activities is discussed in further detail below.

10.6.1 ETL Conventions

ETL conventions are used to ensure as much consistency as possible into the process that each data partner uses when extracting data from their source system and transforming it into the DRN's CDM. There are a number of reasons why this is necessary. The first is that there may not be a 1:1 mapping between the source system and target model. For instance, there may be a concept/domain in the DRN CDM that relates to a patient's smoking status. The allowable values in the CDM are *Heavy Smoker*, *Light Smoker*, and *Smoker, status unknown*. The possible values in the partner's source system, however, are limited to *Current Smoker* and *Never Smoker*. A network can use the ETL conventions document to help provide guidance. In some cases, the decision may be to utilize a specific mapping. In others, partners may be asked to look for additional sources that might help provide more information. The other major area where ETL conventions can benefit a network is when partners can choose between several possible options. For instance, a medication order could be represented using one of several terminologies, and as a brand name or a generic. The conventions will help ensure that a consistent set of decisions is reached across all partners.

10.6.2 Data Quality Assessments

The overall purpose of data characterization routines is to help a network determine whether a given partner's data are suitable for analysis (Brown et al. 2013; Kahn et al. 2012, 2015; Raebel et al. 2014). Suitability may vary based on study type, so networks often deploy assessments with different levels of rigor. These routines are used to look at the variation of data across partners and to help identify potential issues with a partner's ETL procedures, such as mapping errors or data loss. As an example of variation, there may be a check that looks at the average drug exposures (e.g., medication orders, administrations, and/or fulfillments) per person across the network. If most partners have data that range from 30 to 50 exposures/person, but another has ~80 exposures/person, that might indicate that there is an issue with the data. It is possible that the patient population or practice patterns at that partner can explain why their data is an outlier, but it provides rationale for further investigation. Other checks may look at the rate of events over time, which can be useful in identifying whether there was ever an addition or drop of a data feed.

10.6.3 Expanding the Common Data Model

Deciding what to include in a network's CDM is usually based on a number of competing factors. They include the scientific or analytic utility of a given data domain, its availability across partners, feasibility of access, the ability to standardize and the resources required to obtain it. As the number and heterogeneity of partners in a network grows, the sophistication of the model generally decreases, as it can become difficult for every partner to populate each part of the model. Networks may decide to make certain domains optional, which lowers the burden on each partner, but also limits the ability to harness the power of the full network when running a query. In general, all but the most straightforward of studies will require data elements that are outside of the network CDM. To add or expand the data model to include these new elements, networks will need to balance the factors described above in deciding whether to ask all partners to try and obtain these data, whether individual partners should develop study-specific ETL procedures, or if they simply should be obtained via chart abstraction or prospective data collection.

10.7 Conclusion

Learning networks and distributed networks are key components of a national LHS. Learning networks support an array of improvement and care management activities, providing investigators with near-real-time feedback to help improve care delivery. DRNs, on the other hand, enable reproducible, population-scale observational and comparative effectiveness research. As EHRs become more widespread and the underlying data are more frequently mapped to standard terminologies, it should become easier to create and maintain such networks, though it will always be necessary to ensure that the data are of high quality and are fit for the purpose of specific networks.

Acknowledgements Work described in this publication was supported in part by funding from AHRQ (R01 HS020024 and R01 HS022974), by PCORI (CDRN-1306-01556, CDRN-1306-04608, PPRN-1306-01754), by NIH/NCATS (UL1 RR025758S, UL1TR000017/UL1RR026314) and by the participating centers of the ImproveCareNow Network (www.improvecarenow.org). Its contents are solely the responsibility of the author and do not necessarily represent the views or opinions of AHRQ, PCORI, NIH, or the ImproveCareNow Network.

References

Bevans M, Ross A, Cella D. Patient-Reported Outcomes Measurement Information System (PROMIS): efficient, standardized tools to measure self-reported health and quality of life. Nurs Outlook. 2014;62(5):339–45. doi:10.1016/j.outlook.2014.05.009.

Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med. 2010;363(6):501–4.

Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness: the chronic care model, Part 2. JAMA, 2002;288(15):1909–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12377092.

Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, … Trister AD. The mPower study, Parkinson disease mobile data collected using ResearchKit. Scientific Data 2016;3:160011. doi:10.1038/sdata.2016.11.

Brown J, Holmes J, Maro J, Syat B, Lane K, Lazarus R, Platt R. Design specifications for network prototype and cooperative to conduct population-based studies and safety surveillance. Effective Health Care Research Report No. 13 (Prepared by the DEcIDE Centers at the HMO Research Network Center for Education and Research on Therapeutics and the University of Pennsylvania Under Contract No. HHSA29020050033I T05). 2009. Retrieved from www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care. 2010;48(6 Suppl):S45–51. doi:10.1097/MLR.0b013e3181d9919f.

Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care. 2013;51(8 Suppl 3):S22–9. doi:10.1097/MLR.0b013e31829b1e2c.

Buetow KH, Niederhuber J. Infrastructure for a learning health care system: CaBIG. Health Aff (Millwood). 2009;28(3):923–4; author reply 924–925. doi:10.1377/hlthaff.28.3.923-a.

Chung AE, Sandler RS, Long MD, Ahrens S, Burris JL, Martin CF, … Kappelman MD. Harnessing person-generated health data to accelerate patient-centered outcomes research: the Crohn's and Colitis Foundation of America PCORnet Patient Powered Research Network (CCFA Partners). J Am Med Inform Assoc. 2016. doi:10.1093/jamia/ocv191.

Cleghorn GD, Headrick LA. The PDSA cycle at the core of learning in health professions education. Jt Comm J Qual Improv. 1996;22(3):206–12. Retrieved from<Go to ISI>://MEDLINE:8664953.

Coleman K, Austin BT, Brach C, Wagner EH. Evidence on the chronic care model in the new millennium. Health Aff (Millwood). 2009;28(1):75–85. doi:10.1377/hlthaff.28.1.75.

Crandall W, Kappelman MD, Colletti RB, Leibowitz I, Grunow JE, Ali S, … Margolis P. ImproveCareNow: the development of a pediatric inflammatory bowel disease improvement network. Inflamm Bowel Dis. 2011;17(1):450–7. doi:10.1002/Ibd.21394.

Crandall WV, Margolis PA, Kappelman MD, King EC, Pratt JM, Boyle BM, … Colletti RB. Improved outcomes in a quality improvement collaborative for pediatric inflammatory bowel disease. Pediatrics. 2012;129(4):e1030–41. doi:10.1542/peds.2011-1700.

Devoe JE, Gold R, Cottrell E, Bauer V, Brickman A, Puro J, … Fields S. The ADVANCE network: accelerating data value across a national community health center network. J Am Med Inform Assoc. 2014. doi:10.1136/amiajnl-2014-002744.

Drake TA, Braun J, Marchevsky A, Kohane IS, Fletcher C, Chueh H, … Berman J. A system for sharing routine surgical pathology specimens across institutions: the Shared Pathology Informatics Network. Hum Pathol. 2007;38(8):1212–25. doi:10.1016/j.humpath.2007.01.007.

Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014. doi:10.1136/amiajnl-2014-002747.

Forrest CB, Bevans KB, Pratiwadi R, Moon J, Teneralli RE, Minton JM, Tucker CA. Development of the PROMIS (R) pediatric global health (PGH-7) measure. Qual Life Res. 2014a;23(4):1221–31. doi:10.1007/s11136-013-0581-8

Forrest CB, Margolis P, Seid M, Colletti RB. PEDSnet: how a prototype pediatric learning health system is being expanded into a national network. Health Affairs. 2014b;33(7):1171–7. doi:10.1377/hlthaff.2014.0127.

Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, … Kahn MG. PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc. 2014c. doi:10.1136/amiajnl-2014-002743.

Forrest CB, Tucker CA, Ravens-Sieberer U, Pratiwadi R, Moon J, Teneralli RE, … Bevans KB. Concurrent validity of the PROMIS((R)) pediatric global health measure. Qual Life Res. 2016;25(3):739–51. doi:10.1007/s11136-015-1111-7.

Friedman C, Rigby M. Conceptualising and creating a global learning health system. Int J Med Inform. 2012. doi:10.1016/j.ijmedinf.2012.05.010.

Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;2(57):57cm29. doi:10.1126/scitranslmed.3001456.

Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, … Van Houweling D. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J Am Med Inform Assoc. 2015;22(1):43–50. doi:10.1136/amiajnl-2014-002977.

Gershon R, Rothrock NE, Hanrahan RT, Jansky LJ, Harniss M, Riley W. The development of a clinical outcomes survey research application: assessment center. Qual Life Res. 2010;19(5):677–85. doi:10.1007/s11136-010-9634-4.

Greene SM, Reid RJ, Larson EB. Implementing the learning health system: from concept to action. Ann Intern Med. 2012;157(3):207–10. doi:10.7326/0003-4819-157-3-201208070-00012.

Grossman C, Goolsby WA, Olsen LA, McGinnis JM. Engineering a learning healthcare system: a look at the future: workshop summary. Washington, DC: National Academies Press; 2011.

Grossmann C, Powers B, McGinnis JM, editors. Digital infrastructure for the learning health system: the foundation for continuous improvement in health and health care: workshop series summary. Washington, DC: National Academies Press; 2011.

Health Level 7. FHIR Specification Home Page. 2015. Retrieved from http://hl7.org/fhir/.

Hunger SP, Lu X, Devidas M, Camitta BM, Gaynon PS, Winick NJ, … Carroll WL. Improved survival for children and adolescents with acute lymphoblastic leukemia between 1990 and 2005: a report from the children's oncology group. J Clin Oncol JCO. 2012;2011.2037. 8018.

Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012;50(Suppl):S21–9. doi:10.1097/MLR.0b013e318257dd67.

Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, … Zozus MN. Transparent reporting of data quality in distributed data networks. EGEMS (Wash DC). 2015;3(1):1052. doi:10.13063/2327-9214.1052.

Kaushal R, Hripcsak G, Ascheim DD, Bloom T, Campion TR Jr, Caplan AL, … on behalf of the, N.-C. Changing the research landscape: the New York City Clinical Data Research Network. J Am Med Inform Assoc. 2014. doi:10.1136/amiajnl-2014-002764.

Kho AN, Cashy JP, Jackson KL, Pah AR, Goel S, Boehnke J, … Galanter WL. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J Am Med Inform Assoc. 2015;22(5):1072–80. doi:10.1093/jamia/ocv038.

Khurshid A, Nauman E, Carton T, Horswell R. Louisiana clinical data research network: establishing an infrastructure for efficient conduct of clinical research. J Am Med Inform Assoc. 2014. doi:10.1136/amiajnl-2014-002740.

Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc. 2012;19(2):181–5. doi:10.1136/Amiajnl-2011-000492.

Libby AM, Pace WD, Bryan C, Orten Anderson H, Ellis SL, Read Allen R, … Valuck RJ. Comparative effectiveness research in DARTNet primary care practices: point of care data collection on hypogylcemia and over-the-counter abd herbal use among patients diagnosed with diabetes. Medical Care. 2010;48(6, Suppl 1):S39–44.

Loeffler M, Winter A. Editorial: networking clinical research. Methods Inf Med. 2007;46(5):572–3. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17938781.

Mandl KD, Kohane IS, McFadden D, Weber GM, Natter M, Mandel J, … Murphy SN. Scalable collaborative infrastructure for a learning healthcare system (SCILHS): architecture. J Am Med Inform Assoc. 2014. doi:10.1136/amiajnl-2014-002727.

Margolis P, Provost LP, Schoettker PJ, Britto MT. Quality improvement, clinical research, and quality improvement research – opportunities for integration. Pediatr Clin North Am. 2009;56(4):831–41. doi:10.1016/j.pcl.2009.05.008.

Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, Brown JS. Design of a national distributed health data network. Ann Intern Med. 2009;151(5):341–4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19638403.

Marsolo K, Nix J. Workflows for a web-based consent management system with electronic consent (e-consent). Paper presented at the AMIA Summit on Clinical Research Informatics, San Francisco, CA. 2014.

Marsolo K, Margolis PA, Forrest CB, Colletti RB, Hutton JJ. A digital architecture for a network-based learning health system – integrating chronic care management, quality improvement and research. eGEMs (Generating Evidence & Methods to improve patient outcomes), 2015;3(1). doi:http://dx.doi.org/10.13063/2327-9214.1168.

McGinnis JM. Evidence-based medicine – engineering the learning healthcare system. Stud Health Technol Inform. 2010;153:145–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20543243.

McGlynn EA, Lieu TA, Durham ML, Bauck A, Laws R, Go AS, … Kahn MG. Developing a data infrastructure for a learning health system: the PORTAL network. J Am Med Inform Assoc. 2014. doi:10.1136/amiajnl-2014-002746

Moulton G. HMO research network to focus on cancer prevention and control. J Natl Cancer Inst. 1999;91(16):1363. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10451440.

Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124–30. doi:10.1136/jamia.2009.000893.

Namini AH, Berkowicz DA, Kohane IS, Chueh H. A submission model for use in the indexing, searching, and retrieval of distributed pathology case and tissue specimens. Stud Health Technol Inform. 2004;107(Pt 2):1264–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15361017.

Nasseh D, Engel J, Mansmann U, Tretter W, Stausberg J. Matching study to registry data: maintaining data privacy in a study on family based colorectal cancer. Stud Health Technol Inform. 2014;205:808–12.

National Center for Advancing Translational Sciences. ACT Network. 2015. Retrieved from http://act-network.org.

National Institutes of Health. Genetic and rare diseases information center. 2016. Retrieved from https://rarediseases.info.nih.gov/.

Natter MD, Quan J, Ortiz DM, Bousvaros A, Ilowite NT, Inman CJ, … Mandl KD. An i2b2-based, generalizable, open source, self-scaling chronic disease registry. J Am Med Inform Assoc. 2012. doi:10.1136/amiajnl-2012-001042

Office of the National Coordinator for Health Information Technology. Standards & Interoperability (S&I) Framework – Structured Data Capture Initiative. Retrieved from http://wiki.siframework.org/Structured+Data+Capture+Initiative.

Ohno-Machado L, Agha Z, Bell DS, Dahm L, Day ME, Doctor JN, … the p, S. t. pSCANNER: patient-centered Scalable National Network for Effectiveness Research. J Am Med Inform Assoc. 2014. doi:10.1136/amiajnl-2014-002751.

Olsen LA, Aisner D, McGinnis JM, editors. The learning healthcare system: workshop summary. Washington, DC: National Academies Press; 2007.

Olsen LA, Saunders RS, McGinnis JM, editors. Patients charting the course: citizen engagement and the learning health system: workshop summary. Washington, DC: National Academies Press; 2011.

Pace WD, Cifuentes M, Valuck RJ, Staton EW, Brandt EC, West DR. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med. 2009a;151(5):338–40. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19638402.

Pace WD, West DR, Valuck RJ, Ciefuentes M, Staton EW. Distributed Ambulatory Research in Therapeutics Network (DARTNet): summary report (Prepared by the University of

Colorado DEcIDE Center Under Contract No. HHSA290200500371 TO2.). 2009b. Retrieved from Rockville, MD: http://www.effectivehealthcare.ahrq.gov/ehc/products/53/151/2009_0728DEcIDE_DARTNet.pdf

PCORnet PPRN Consortium, Daugherty SE, Wahba S, Fleurence R. Patient-powered research networks: building capacity for conducting patient-centered clinical outcomes research. J Am Med Inform Assoc. 2014. doi:10.1136/amiajnl-2014-002758.

PopMedNet. PopMedNet. 2011. Retrieved from http://www.popmednet.org/Home.aspx.

Raebel MA, Haynes K, Woodworth TS, Saylor G, Cavagnaro E, Coughlin KO, … Brown JS. Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel. Pharmacoepidemiol Drug Saf. 2014;23(6):609–18. doi:10.1002/pds.3580.

Randell RL, Long MD, Cook SF, Wrennall CE, Chen W, Martin CF, … Kappelman MD. Validation of an internet-based cohort of inflammatory bowel disease (CCFA partners). Inflamm Bowel Dis. 2014;20(3):541–4. doi:10.1097/01.MIB.0000441348.32570.34.

Raths D. Trend: standards development. Catching FHIR. A new HL7 draft standard may boost web services development in healthcare. Healthc Inform. 2014;31(2):13, 16. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24812983.

Rosenbloom ST, Harris P, Pulley J, Basford M, Grant J, Dubuisson A, Rothman RL. The mid-south clinical data research network. J Am Med Inform Assoc. 2014. doi:10.1136/amiajnl-2014-002745.

Ross JA, Severson RK, Robison LL, Pollock BH, Neglia JP, Woods WG, Hammond GD. Pediatric cancer in the United States. A preliminary report of a collaborative study of the childrens cancer group and the pediatric oncology group. Cancer. 1993;71(10 Suppl):3415–21. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8490892.

Schmidlin K, Clough-Gorr KM, Spoerri A. Privacy preserving probabilistic record linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. BMC Med Res Methodol. 2015;15:46. doi:10.1186/s12874-015-0038-6.

Selby JV, Beal AC, Frank L. The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda. JAMA. 2012;307(15):1583–4. doi:10.1001/jama.2012.500.

Selby JV, Forsythe L, Sox HC. Stakeholder-driven comparative effectiveness research: an update from PCORI. JAMA. 2015;314(21):2235–6. doi:10.1001/jama.2015.15139.

Shah A, Stewart AK, Kolacevski A, Michels D, Miller R. Building a rapid learning health care system for oncology: why CancerLinQ collects identifiable health information to achieve its vision. J Clin Oncol. 2016;34(7):756–63. doi:10.1200/JCO.2015.65.0598.

Sledge GW Jr., Miller RS, Hauser R. CancerLinQ and the future of cancer care. Am Soc Clin Oncol Educ Book. 2013;430–4. doi:10.1200/EdBook_AM.2013.33.430.

Standards & Interoperability (S&I) Framework – Health eDecisions Homepage. 2013. Retrieved from http://wiki.siframework.org/Health+eDecisions+Homepage.

Steiner JF, Paolino AR, Thompson EE, Larson EB. Sustaining Research Networks: the twenty-year experience of the HMO research network. EGEMS (Wash DC). 2014;2(2):1067. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25848605.

The caBIG Strategic Planning Workspace. The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community. Stud Health Technol Inform. 2007;129(Pt 1):330–334. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17911733.

Toh S, Platt R, Steiner JF, Brown JS. Comparative-effectiveness research in distributed health data networks. Clin Pharmacol Ther. 2011;90(6):883–7. doi:10.1038/clpt.2011.236.

Vogt TM, Elston-Lafata J, Tolsma D, Greene SM. The role of research in integrated healthcare systems: the HMO Research Network. Am J Manag Care. 2004;10(9):643–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15515997.

Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR. The greater plains collaborative: a PCORnet clinical research data network. J Am Med Inform Assoc. 2014. doi:10.1136/amiajnl-2014-002756.

Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16(5):624–30. doi:M3191 [pii].

Chapter 11
Natural Language Processing – Overview and History

Brian Connolly, Timothy Miller, Yizhao Ni, Kevin B. Cohen, Guergana Savova, Judith W. Dexheimer, and John Pestian

Abstract In this chapter, we introduce the topic of Natural Language Processing (NLP) in the clinical domain. NLP has shown increasing promise in tasks ranging from the assembly of patient cohorts to the identification of mental disorders. The chapter begins with a discussion of the necessity of NLP for analyzing

B. Connolly, Ph.D. (✉)
Department of Pediatrics, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Avenue, Cincinnati, OH 45229-3039, USA
e-mail: brian.connolly@cchmc.org

T. Miller, Ph.D.
Department of Pediatrics, Harvard Medical School, Boston Children's Hospital, 300 Longwood Avenue Enders 138, Boston, MA 02115, USA
e-mail: timothy.miller@childrens.harvard.edu

Y. Ni, Ph.D.
Department of Pediatrics and Biomedical Informatics, Division of Biomedical Informatics, Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Avenue, ML-7024, Cincinnati, OH 45229-3039, USA

K.B. Cohen, Ph.D.
University of Colorado School of Medicine, 13001 E 17th Pl, RC-1 S. Room L18-6102, Aurora, CO 80045, USA
e-mail: Kevin.Cohen@gmail.com

G. Savova, Ph.D.
Children's Hospital Boston and Harvard Medical School, 300 Longwood Avenue, Enders 138, Boston, MA 02115, USA
e-mail: Guergana.Savova@childrens.harvard.edu

J.W. Dexheimer, Ph.D.
Departments of Pediatrics and Biomedical Informatics, Divisions of Emergency Medicine and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Ave, ML-2008, Cincinnati, OH 45229, USA
e-mail: judith.dexheimer@cchmc.org

J. Pestian, Ph.D., M.B.A.
Department of Pediatrics and Biomedical Informatics, Division of Emergency Medicine, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Ave, ML-2008, Cincinnati, OH 45229-3039, USA
e-mail: john.pestian@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_11

EHRs. Subsequent sections then place clinical NLP research in a wider historical context by reviewing various approaches to NLP over time. The focus then turns to available NLP-related data resources and the methods of generating such resources. The actual development of NLP systems and their evaluation are then examined. The chapter concludes by describing current and future challenges in clinical NLP.

Keywords Annotation • Evaluation • Gold standard • Machine learning • Natural language processing • Performance measures • Text analysis • Statistical Measures

11.1 Introduction to Clinical Natural Language Processing

Natural language processing (NLP) describes the application of computational methods to datasets of human languages, often for the purpose of extracting information into a format that makes it more suitable for use in downstream applications. In the biomedical domain, typical inputs include biomedical journal texts and clinically-generated text. This chapter will focus on NLP with a few example use cases and will describe the types of information and representations that may be extracted from biomedical texts. The next chapter will focus on specifics of some downstream applications of biomedical NLP.

11.2 Motivation

The Electronic Health Record (EHR) is a vast repository of information about a patient, much of which is in free text written by various health professionals during the course of care. The importance of information included only in the narrative clinical text of the EHR is gaining increasing recognition as a critical component of computerized decision support, quality improvement, and patient safety (Meystre et al. 2008). The field of NLP which focuses on the clinical narrative is dubbed clinical NLP. As such, NLP is an enabling technology for accessing the vast information residing in the EHR notes (Jha 2011). For example, NLP could extract information from clinical free-text to populate decision rules or represent clinical knowledge and procedures in a standardized format (Demner-Fushman et al. 2009). Approximately 30–50 % of data elements useful for quality improvement are available only in the text of the EHR (Hicks 2003). In addition, patient safety and clinical research could also benefit from information stored in text that is not available in either structured EHR entries or administrative data (Melton and Hripcsak 2005; Murff et al. 2011; Zhan and Miller 2003).

Further, NLP technologies based on patient-generated data (social media, patient interviews, written word, etc.) have shown increasing promise for decision support tools that aid in the diagnosis and treatment of neuropsychiatric and developmental disorders.

Four examples of the use of NLP to support translational research are provided here (and revisited in detail in Chap. 12) to help the reader understand the power of the methodology and to set the stage to learn more about the basic principles of NLP (Denny et al. 2010; Kho et al. 2011; Kohane 2011):

- Deidentification of text in EHRs: NLP permits the automated removal of personal health information from vast numbers of records in the electronic patient data warehouse, now maintained by many medical institutions to support clinical care delivery. Deidentification permits an investigator to review the EHRs for large numbers of patients prior to receiving institutional review board approval of specific projects requiring consent of patients to review personal health information.
- Phenotyping in support of Genomics Research: Discrete data in EHRs typically includes coding of diagnoses by International Classification of Disease codes (ICD-9, ICD-10) and coding of procedures performed by Current Procedural Terminology (CPT) codes (CPT). Assembling cohorts of patients with a specific diagnosis (e.g., diabetes or appendicitis) for genomic studies requires highly reliable phenotyping. Coding in medical records in the United States is generally for billing purposes and not for research. For this reason claims data are not sufficiently reliable to support genomics research. Applying NLP methods to clinical notes to confirm diagnoses, greatly increases the accuracy of phenotypes and facilitates genetic and genomic studies (Denny et al. 2010; Kho et al. 2011; Kohane 2011). The successes of eMERGE and PGRN (eMERGE; National Institutes of Health and National Institute of General Medical Sciences) networks clearly demonstrated the essential role of NLP (Giacomini et al. 2007; McCarty et al. 2011).
- Pharmacovigilance and novel discoveries as demonstrated by the Vioxx work (Brownstein et al. 2007). Comparative effectiveness and studies as discussed in Chap. 10, and epidemiology studies (Brownstein et al. 2009; Health Map 2006). Through an automated process, updating 24/7/365, HealthMap monitors, organizes, integrates, filters, visualizes and disseminates online information about emerging diseases in nine languages, facilitating early detection of global public health threats.
- In the neuropsychiatric and developmental domain, patient-generated data can provide important 'primary source' information about a patient's condition. For instance, the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al. 2001), which provides psychological insights from a person's words, has been used to predict post-bereavement improvements in mental and physical health (Pennebaker et al. 1997) and adjustment to cancer (Demner-Fushman et al. 2009; Owen et al. 2006). As will be discussed in Chap. 12, suicide notes have provided insight into the language of suicide, while the language in psychiatric interviews and social media has been used to identify suicidal risk (Desmet 2014; Gomez 2014; Huang et al. 2007; Jashinsky et al. 2014; T. Li et al. 2013b; Matykiewicz et al. 2009; Pestian et al. 2008; Thompson et al. 2014; Zhang et al. 2014).

11.3 Background and History

NLP research effectively falls into two major eras reflecting different paradigms. The first era focused on rule-based approaches and theoretical foundations, following progress in theoretical linguistics and using computational reasoning systems to implement formal logic believed to encode linguistic knowledge. The following era (which we are still in) focuses on data-driven approaches, making use of large available datasets (such as the World Wide Web or EHRs) as well as gains in computing power to build sophisticated statistical models of language.

11.4 Rule- and Logic-Based Methods for Natural Language Processing

Initial forays into NLP technology used some of the first computers to implement rule-based methods that underestimated and oversimplified the task of language understanding. This first phase started in the late 1940s and continued until the late 1960s. Initially, this phase was dominated by optimism and great expectations for impending success in Machine Translation (MT). Automatic MT focused on Russian to English, driven by the needs of the military and intelligence gathering communities. Translation efforts were based on word-for-word processing and dictionary lookups. Syntactic and semantic ambiguities were handled by analyzing local context and utilizing elaborate hand crafted rules. The ALPAC Report in 1966 criticized these MT efforts (National Research Council 1966), but also provided recommendations to spend more on "*computational linguistics as a part of linguistics– studies of parsing, sentence generation, structure, semantics, statistics, and quantitative linguistic matters, including experiments in translation*" (National Research Council 1966). All of these components became major lines of research within the computational linguistics community, which led to breakthroughs in computational text processing. Influential work from this era includes Chomsky's work (1965) launching formal language theory (including the context-free grammar) and Shannon's ideas of noisy channel and entropy (Shannon and Weaver 1949), which led to probabilistic algorithms for speech and language processing.

The late 1960s and 1970s were dominated by Artificial Intelligence (AI) researchers and question answering systems focusing on narrowly defined topics. Partly because of the lessons learned during the previous decades and because transformational grammar (the dominant linguistic theory of the time) was considered ill-suited for computational tasks, the potential contribution of the linguistics field was mostly ignored. A number of methods were explored in the areas of syntactic parsing, mainly incorporating rule-based approaches.

Subsequent rule-based NLP research (between the late 1970s and 1980s) was dominated by developments in computational grammar theory (e.g. functional, categorical, and generalized phrase structure grammars) and linking them with AI

logics for meaning and knowledge representation. (D)ARPA funded speech recognition and message understanding conferences promoted rigorous performance evaluation the first time in NLP.

11.5 The Statistical Revolution

The (D)ARPA initiated focus on evaluation (1990s to 2000s) was a major factor in the ascent of statistical methods in the general NLP field. Large-scale practical tasks were successfully tackled and corresponding robust systems were developed (e.g. Internet search engines). In addition, government funded evaluation efforts (e.g. for robustness and portability in public NLP challenges), and development of linguistic resources (e.g. British National Corpus (Aston and Burnard 1998), WordNet (WordNet), Penn Treebank (Penn Treebank Project), PropBank (PropBank Project), etc.) and tools (e.g. part of speech taggers, chunkers) dominated this period.

The most important recent trend has been working with big data (Manyika et al. 2011) (for example the text data on the entire internet), and computationally powerful numerical methods for representing domain knowledge, tackling each discrete NLP task – syntactic parsing, predicate-argument structure, entity mention discovery, relation identification, temporality, overall text cohesiveness, text generation, summarization, machine translation. Each of these tasks has grown into a major line of computational linguistics research, the output of which has matured to a level where it can be successfully incorporated into applications (see Chap. 12 for more details on applications). There are many excellent resources for the reader interested in more history of the field of natural language processing (Bright 1992; Jones 2001; Jurafsky and Martin 2000; Zampolli et al. 1994).

11.6 Data-Driven vs. Knowledge-Driven Approaches

The distinction between rule-based and data-driven methods is still relevant today. Most systems in production are likely to be hybrids, using rules and statistical methods, where each has its strength. Rule-based methods excel at interpretability, often important in the medical domain, while suffering from brittleness and requiring ongoing maintenance. Statistical methods can be more robust to variation in input, but often require large hand-labeled datasets to learn reliably.

A real example of a supervised machine learning (data-driven) (ML) approach versus a rule-based approach is presented in an Acute Lung Injury (ALI) chest X-ray classification paper (Solti et al. 2009). In the rule-based approach the classification keywords are predefined by domain experts (Pulmonary Medicine physicians). In the supervised ML (data-driven) approach the algorithm "learns" the important keywords from the annotated corpus and assigns weights of importance to the identified keywords on its own. For a highly accurate supervised ML

algorithm, a well-annotated large training corpus is vital. In the ALI example, the annotated corpus consists of a set of chest X-ray reports that were classified by physicians as consistent or not with the diagnosis of ALI. When a sufficiently large corpus is annotated with proper annotation methodology then the supervised automated learning is robust and usually provides better results than the rule-based system.

The biomedical NLP domain has lagged behind general-domain NLP in its adoption of statistical methods, for a variety of reasons. One likely contributing reason is that data access is more privileged for clinical tasks – obtaining millions of words of general-domain text requires nothing more than an internet connection, while obtaining millions of words of medical text is quite a bit more involved. Even where shared datasets exist, the bureaucratic requirements in obtaining these datasets may limit their adoption to the truly driven. However, a more positive reason for the prevalence of rule-based and knowledge-based methods in the clinical domain is the availability of high-quality resources. The Unified Medical Language System (UMLS (Bodenreider and McCray 2003; UMLS [Internet])), provides an excellent knowledge resource that, among other things, includes synonym lists for hundreds of thousands of clinical terms. With such a resource available and maintained by the National Library of Medicine, statistical methods for entity recognition – the mapping of a text span to a concept – are not as necessary. It is also likely that, given the high quality of knowledge-based resources of the UMLS, funding agencies were slow to recognize the value of building large corpora of clinical data for training NLP methods.

Recognizing the importance of creating linguistic datasets for the clinical domain, the National Institutes of Health agencies have funded a variety of clinical NLP initiatives to build clinical corpora with linguistic and clinical annotations (Allbright et al. 2012; Pradhan et al. 2014; Styler et al. 2014).

11.7 Data Resources for Clinical Natural Language Processing

The medical domain comprises a variety of sources of natural language data. As much as 30–50 % of clinical data are only found in natural language notes (Hicks 2003). Accessing and using the information contained in those documents is the goal of medical NLP systems. To develop as well as to evaluate the robustness of such systems, high quality gold standards are required. That is, sets of manually labeled instances that are relevant to the specific NLP tasks must be created. For information extraction tasks, creating such gold standards involves looking at text corpora and labeling named entities (relevant text phrases) with the proper category (e.g. *disease*, *symptom*, *medication*, etc.) and/or the relations between them (e.g. *location of relation*). Such a process is called gold standard annotation creation, which is usually accomplished by manual annotations.

In the general domain, the creation of the Penn Treebank (Marcus et al. 1993) and the word sense-annotated SEMCOR (Fellbaum et al. 1998) showed how even limited amounts of annotated data can result in major improvements in complex natural language understanding systems. Since then, there have been many other successful efforts including the Automatic Content Extraction (ACE) annotations (Named Entity tags, nominal entity tags, coreference, semantic relations and events); semantic annotations, such as more coarse grained sense tags (Palmer et al. 2007); semantic role labels as in PropBank (Palmer et al. 2005), NomBank (Meyers et al. 2004), and FrameNet (Baker et al. 1998); and pragmatic annotations, such as coreference (Poesio 2004; Poesio and Vieira 1998), temporal relations as in TimeBank (Pustejovsky et al. 2003; Pustejovsky and Stubbs 2012), the Opinion corpus (Wiebe et al. 2005), and the Penn Discourse Treebank (Prasad et al. 2008) to name just a few. Many of the core NLP components listed in Sect. 11.3 are trained and evaluated on these annotations. For details on general annotations, consult the excellent expositions in Palmer and Xue (2010) and Pustejovsky and Stubbs (2012).

11.8 Data Resources in the Clinical Domain

Several studies have built corpora of clinical notes annotated with medical information to train and evaluate the performance of NLP systems. They are often focused on a specific entity type or on a specific topic. Ogren et al. (2008) built a gold standard corpus of clinical notes annotated for disorders in order to measure the performance of a NLP system that processes a repository of clinical notes and identifies mentions of disorders as well as the context (e.g., "current", "history of", etc.) and status (e.g., "confirmed", "negated", etc.) of those mentions. Deleger et al. (2014) created a gold standard set of protected health information (PHI) annotation for clinical notes. By releasing the annotated corpus with Data User Agreement, they facilitate the development of new machine learning models for de-identification research. Uzuner et al. (2010) evaluated the performance of several medication extraction systems against a corpus manually annotated for medication names and related information such as dosage and frequency. This corpus provided a benchmark for accurately comparing systems. South et al. (2009) developed a manually annotated corpus for a use case of identifying phenotypic information for inflammatory bowel disease. A few studies reported annotation effort on a larger scale, including multiple entities as well as relations between them, although they are often limited in size and/or to specific types of clinical documents. The annotated corpus of the Clinical E-Science Framework (the CLEF corpus) is one of the richest semantically annotated resources (Roberts et al. 2009). It includes a variety of named entities, relationships, coreference, and temporal information, and gathers data from the whole patient record, although only a few hundred documents were annotated. The authors demonstrated its use in developing and evaluating an information extraction system that identifies medical entities such as conditions, drugs, devices and interventions, as well as a variety of relationships between those entities

(e.g. has_finding, has_location, etc.). The corpus from the 2010 i2b2 Informatics for Integrating Biology & the Bedside (i2b2), NLP shared task (Uzuner et al. 2011) is also annotated for several entities (treatments, problems, and tests), modifiers (e.g., negation and uncertainty), and relationships, but is limited to discharge summaries. It has been used as a benchmark to evaluate and compare a number of NLP systems extracting medical problems, treatments and tests along with their modifiers and relationships. Chapman and Dowling (2006), Chapman et al. (2008) developed an extensive annotation schema to capture medical conditions and focused their annotation effort on emergency department reports. Their goal was to provide standard guidelines for manually annotating variables relevant for the indexing of clinical conditions (such as symptoms, diagnoses, radiological findings, etc.) in order to create strong reference corpora to evaluate automated NLP-based indexing applications (i.e., applications that store the output in easily searchable format). Automatically indexing clinical conditions from emergency reports can be useful for many purposes, including identifying patients eligible for clinical trials.

Most of the annotation effort in the clinical domain has been focused on notes from the EHR. Other types of medical documents, such as clinical trial announcements or FDA drug labels, have been much less frequently annotated. Clinical trial announcements are starting to draw interest and have been annotated on a few occasions: for a preliminary study on using Amazon Mechanical Turk (Yetisgen-Yildiz et al. 2010) and for detecting temporal constraints in a sample of 100 eligibility criteria sentences (Luo et al. 2011). On-going efforts are investigating large-scale annotation of both clinical trial announcements and FDA drug labels (Q. Li et al. 2013a). These corpora are being annotated for a variety of medical concepts, including medications, symptoms, and diagnoses. Intended applications are the building of automated NLP systems to identify children eligible for clinical trials and to mine adverse drug events.

Clinical texts have also annotated for critical linguistic layers. For instance, Pakhomov et al. (2006) have annotated a corpus of clinical notes for part-of-speech information in order to adapt part-of-speech taggers to the medical domain, as part-of-speech tagging is a pre-processing task essential to many information extraction systems.

A very recent trend in the clinical NLP is to create sharable, large-scale, multi-layered, extendable annotations similar in style to those available in the general domain as outlined in the very beginning of Sect. 11.4. In the past several years in the medical informatics community, several complementary initiatives have started leveraging large de-identified clinical corpora and establishing schemas and guidelines for their annotations at the syntactic, semantic, and discourse levels. The goal of all these efforts is to provide the research community with shared annotated corpora to allow for the development and testing of novel clinical NLP methods and tools to meet the aspirations of sophisticated natural language understanding systems that can be used in a variety of translational and healthcare research use cases. Special care was given to ensure that the initiatives are complementary and aligned, so that the resulting resources can be merged transparently whenever possible. The general principles are to (1) ensure that the annotated resources are available to the

research community through Data Use Agreements with the contributing institutions, (2) establish smooth mechanisms of delivery and maintenance of the annotation (3) create a set of annotations that are for general NLP purpose, yet specific enough to enable the development of meaningful clinical applications, (4) adhere to standards where such exist, (5) rely on and normalize to clinical terminologies and ontological knowledge when annotating the clinical core concepts in the corpus (e.g., SNOMED-CT, RxNORM); (6) ensure compatibility with existing general-domain annotated resources, by utilizing community-adopted conventions whenever possible (e.g., the PTB part-of-speech tagset); (7) rely on existing schemas and guidelines in the general domain and in the clinical domain whenever possible (e.g., syntactic chunking and full dependency parses); (8) for the clinical-specific annotation, design schemas with modular extensions, which act as additional layers on top of existing, established general-domain schema. In the *Shared Annotated Clinical Resource project* (Pradhan et al. 2014; ShaRe), 500,000 tokens of clinical narratives from MIMIC corpus are annotated with linguistic information (POS tags and phrasal chunks) and full semantic parses of disorder mentions (e.g., core clinical concept normalized to terminology, along with modifiers such as anatomical location, temporality, severity, etc.). The gold standard annotations developed under the Multi-source integrated platform for answering clinical questions (*MiPACQ* (Allbright et al. 2012)) amount to approximately 200,000 tokens of clinical narrative, Medpedia content, and clinical questions from Ely et al (2005). In addition to constituent syntactic parses and semantic roles, the corpus includes UMLS entities, coreference, UMLS relations and answer types. In the Temporal Histories of Your Medical Events (*THYME* (Styler et al. 2014; THYME)), an additional layer is added – that of temporal relations between clinical events – based on the ISO TimeML specifications (TimeML) to a corpus of 200,000 tokens of clinical narrative. Under the *Strategic Health Advanced Research Project Area 4* (SHARPn.org), 500,000 tokens of clinical narrative spread across specialties, patients, notes types, and two sites are annotated. Layers of annotations consist of constituent parses, dependency parses, semantic role labels, UMLS entities, coreference, UMLS relations, and five templates based on the Clinical Element Models (CEMs).

11.9 The Process of Creating Gold Standard Annotations

The main components of the process of creating gold standard annotations are (1) determining a representative textual dataset, (2) developing annotation guidelines, schemas and pilot annotations, (3) training the annotators, (4) single annotations by two annotators followed by adjudication of disagreements, (5) tracking quality of annotations through inter-annotator agreement. Each component is discussed below.

11.10 Data

The data, which becomes the basis for creating the gold annotations,can be (1) written text from the clinical narrative, (2) health-related social media text, (3) health-related newswire text, (4) health-related speech. In the previous section, we introduced gold resources where the data source was already represented by text. Below we describe speech as a data source and how it is transformed into a workable NLP-friendly stream.

Processing speech requires a different level of capture, where all distinct speech characteristics are faithfully recorded, including speech-specific phenomena such as filled pauses (e.g., *uh* and *um*), silences, re-starts, gestures, and other non-verbal events. This data source is very different from the other types of health-related data. Its capture is accomplished through a transcription process with clearly defined, but not excessive guidelines, of how to transcribe speech characteristics. For instance, full phonetic transcriptions might not only be time-consuming, but may hinder those who want to carry out simple analyses, which only rely on the occurrences of certain words without regard to their pronunciation. However, in certain tasks, the important features that determine diagnosis or treatment are paralinguistic features, such as number and duration of pauses, pace of speech, laughter, sighs, and filled pauses (e.g., (Pestian et al. 2016)). Also, a patient's mental or physical state may be evidenced as much through words as by simple gestures (e.g., head nods, head turns, etc.(Birdwhistell 2010)). In cases where NLP is used to understand child development or aphasia, phonetics, contractions, reduction processes (e.g., the de-emphasis of word endings), or even invented words may require annotation (Ratner et al. 1996; Young 1987). Further, the transcription format itself may need to be sufficiently flexible to capture environment, sentence structure, general comments, or even the truthfulness of the patient's statements (MacWhinney 2000).

Given then that any conversation can be described in infinite detail, transcription guidelines are usually determined within four constraints: (1) transcription time, (2) judgments to be made by the transcriptionist or interviewer, (3) transcriptionist experience, and (4) the envisioned NLP task/analyses. The **transcription time** can vary dramatically depending on what is required. A professional transcriptionist typically requires three hours to transcribe one hour of audio. However, the inclusion of time-stamps and other annotations may increase the transcription time by a factor of 10. **Judgments** range from determining whether a sentence is continued in the midst of an interruption to noting important changes in the interview environment. **Transcriptionist Experience** is important to ensure the transcriptionist is able to efficiently follow complex transcription guidelines and emotionally handle the content. The transcriptions should allow a diverse set of **NLP analyses** with little pre-processing of data. For instance, the transcription of the utterance "beginin'" as "beginin[g]" easily allows NLP analyses that both include or exclude reduction processes by removing characters in brackets or by ignoring the brackets entirely.

Different NLP projects have chosen to leverage these constraints differently. For instance, the goal of the Fisher Corpus (Cieri et al. 2004) was to create a large corpus that could be used to develop automated speech recognition tools. This involved transcribing over 2000 h of English language telephone conversations over the course of a single year. The conversations therefore had to be transcribed quickly with time-stamps in a format that could be analyzed using NLP. The Quick Transcription Specification guidelines (Cieri et al. 2004) were then employed, which outlined minimal language transcription requirements (no special effort to provide capitalization, punctuation or special indicators of background noise, mispronunciations, nonce words or the like), and automated methods of time-stamping each word. The corpus was then "NLP-friendly", as the transcription requirements were strict, but minimal.

In contrast, the CHAT transcription format, used by CHILDES (MacWhinney 2000), allowed a maximal number of requirements, but provided software to facilitate the transcription task. The CHAT format allowed for detailed transcriptions with time-stamps, and allowed for the transcription of a wide range of communication – even sign language. Faced with the guidelines' complexity, the CHILDES project developed a program called Computerized Language Analysis (CLAN(MacWhinney 2000)), which allowed utterances to be easily time-stamped and transcribed.

Whether the focus is on rich detailed annotations for scientific purposes, quickly generating usable data for engineering purposes, or somewhere in between, it is important to monitor transcription quality. There are no "standard" measures, as the definitions of transcription error are dependent on the project. For instance, the American Association for Medical Transcriptionists (AAMT) defines acceptable accuracy based on "levels" of error (critical, major, minor, and dictation flaws), which are determined by the errors' effect on patient safety (AAMT 1994, 2005). Although what defines an error may change depending on the NLP task, it remains useful to structure the severity of errors in a similar manner.

11.11 Annotators

Manually annotating medical documents can require different levels of skill, depending on the complexity of the annotation task. When annotating non-medical entities, annotators are not required to have a medical background. For instance, annotating Protected Health Information elements such names and dates for evaluating a de-identification system can be performed perfectly well by non-clinicians. Medical knowledge is most useful when annotating medical entities and/or their relations (e.g. disorders, medications, location of relation, etc.). However, with proper training, non-clinicians can accomplish high quality annotation for this type of task. Chapman et al. (2008) used both clinician and non-clinician annotators to annotate medical conditions in emergency department reports and found that the annotation by both was high quality but non-clinicians needed more extensive

training. For linguistic level annotations such as parsers, temporality, coreference (see the description of the THYME, MiPACQ, SHARP and ShARe corpora), experts with linguistic background are needed.

11.12 Process

A typical annotation process involves using pairs of annotators to independently perform single annotations on the same text followed by an adjudication phase to resolve disagreements. Inter-annotator agreement is tracked to evaluate the reliability of the gold standard. Double-annotation of documents allows reducing cases of mislabeling and building stronger gold standards. Annotators rely on guidelines to perform annotation. Guidelines define the annotation schema, i.e. the specific items to be annotated, and provide specific examples to guide the annotation. Ideally guidelines should be non-ambiguous and in a stable form when given to the annotators. In actuality guidelines are often revised or extended with more examples during the annotation process when previously unseen cases arise. It is also good practice to include a pilot phase in an annotation project (Palmer and Xue 2010). This phase is essential to get annotators accustomed to the guidelines and avoid any misinterpretation. Also, the data can present unforeseen cases, which will require a revision of the guidelines.

There are a number of annotation tools that have been used by many projects – Knowtator (Ogren 2006), Anafora (Chen and Styler 2013), BRAT (brat) CLAN (CLAN), TOBI (Silverman et al. 1992), and ELAN (TLA ; Wittenburg et al. 2006) to name just a few. The community is actively working on better and more robust tools and the list becomes quickly outdated.

Reliability of manually created gold standards is usually measured through inter-annotator agreement (IAA). Good IAA is important because, if agreement is low, then it is likely that the task is too difficult or not well defined (Gale et al. 1992). Moreover, it is generally considered an upper bound on the measured automated performance but not necessarily system ceiling (Gale et al. 1992). There are many ways to quantify IAA (Altman 1991; Banerjee et al. 1999; Osborne 2008). In this section, we will discuss three common measures of IAA used in the NLP and medical literature: Cohen's kappa coefficient, F-measure, and Krippendorff's alpha.

Cohen's kappa coefficient and F-measure are evaluated in tasks involving only two annotators. Cohen's kappa coefficient (k) (Cohen 1960) is defined in terms of the observed fraction of annotations that agree (A_o) and fraction of annotations that are expected to agree by chance (A_e) (Fig. 11.1):

$$\kappa = \frac{A_o - A_e}{1 - A_e}.$$

Fig. 11.1

Since the fractions of agreement (A_x's) have common denominators, the numerator can be re-interpreted as the number of observed agreements above the number of agreements that would be expected by chance. Similarly, the denominator can be re-interpreted as the number of agreements, if there were perfect agreement (i.e. the number of items) above the number of agreements that would be expected by chance. This suggests agreement above what would be expected from chance, $\kappa = 0$ suggests agreement by chance, and $\kappa < 0$ suggests disagreement is systematic and below what would be expected by chance. There is no commonly accepted value indicating "good" agreement, and, as with the performance measures described below, statistically significant deviations from random chance agreement ($\kappa = 0$) are usually determined through bootstrapping methods.

Cohen's kappa is often the standard measure agreement between two annotators. However, kappa is not the most appropriate measure for annotation of named entities in textual documents (Tomanek and Hahn 2009), because it requires the numeration of negative cases, which is unknown for named entities; named entities are sequences of words, and there is no pre-existing fixed number of items to consider when annotating a text. Proposed solutions include considering individual tokens as the items to be marked to compute a "token-level" kappa (Grouin et al. 2011), considering only noun phrases, considering all possible n-grams in a text (sequences of n tokens) or considering only items annotated by one or two of the annotators (Geisser 1975). However, none of those options are fully accurate (Geisser 1975), which is why the F-measure, which does not require the number of negative cases, is usually used to measure IAA for these type of annotation tasks. F-measure is the harmonic mean of precision (positive predictive value) and recall (sensitivity), which is defined as (Fig. 11.2)

$$F_\beta = \frac{(1+\beta^2)Precision * Recall}{\beta^2 * Precision + Recall}$$

Fig. 11.2

where b is an adjustable parameter. The introduction of the b parameter allows the measure to be more or less influenced by the precision measure in the calculation. When computing F_β, the annotations of one annotator are considered as "truth" while the annotations of the other annotator are considered as the predictions of a system. Switching the annotators' roles gives the same F-measure value so it does not matter who plays the role of the reference (Hripcsak and Rothschild 2005; Tomanek and Hahn 2009).

While the value of b is arbitrary, it is usually set to unity, allowing the precision and recall to be weighted equally. In the case where $b = 1$, F_β reduces to (Fig. 11.3)

$$F_1 = \frac{2*t_p}{2*t_p+f_p+f_n},$$

Fig. 11.3

where t_p is the number of "true positives", f_p is the number of "false positives", and f_n is the number of "false negatives". Expressed in this way, F_1 bears a close resemblance to the equation for accuracy, where the true negatives are replaced by t_p.

The F-measure carries several pros and cons. The b parameter allows a certain degree of customization, and it can be reduced to an accuracy-like measure when $b = 1$. Also, the F-measure ignores true negatives, which is advantageous if the focus of the task is to identify one subject group (in this case, inter-annotator agreements). Further, at least in the NLP literature, there is a general consensus as to what constitutes "good agreement" ($F_1 \geq 80\,\%$). On the other hand, a highly imbalanced data set with an infinite number of imperfectly classified items from the "negative" class drives F_b to 0, and so unlike κ or accuracy, it is important to always compliment the observed F-measure with the F-measure value expected from chance agreement. (Unlike Cohen's kappa, random chance agreement in not folded into the calculation of F_β.) In addition, the F-measure is rarely found in the medical literature. Whereas Cohen's alpha and F-measure can only calculate inter-annotator agreement in cases of two annotators, Krippendorff's alpha (a) (Gwet 2011; Krippendorff 2007) carries no such restriction. In contrast to Cohen's kappa, Krippendorff's alpha is most intuitively parameterized in terms of observed *disagreement* and is generally written as (Fig. 11.4)

$$\alpha = 1 - \frac{D_O}{D_e},$$

Fig. 11.4

where D_o parameterizes the observed disagreement, and D_e is the disagreement expected by random chance. (The analogue between k and a is clear when D_o is set to $1-A_o$, and D_e is set to $1-A_e$) The observed disagreement is calculated by averaging the disagreement over all pairs within each "analysis unit" (i.e., question asked to the annotator), u, and then obtaining the weighted average of differences over all units, where each unit is weighted by the fraction of annotations in that unit.[1]

[1] Mathematically, the average inner-unit disagreement is given by
m_u,the number of annotators for u, $d(x,y)$ is an arbitrarily defined function that quantifies dis-similarity between two values x and y, and a_{jk} is the value assigned to the kth unit by the jth annotator. D_o is then defined as

$$D_u = \frac{1}{m_u} \sum_{i=1,i' \neq i}^{m} \delta(a_{iu}, a_{i'u})$$

where there are n pairable values over m_u annotators and N analysis units. Note each D_u is weighted by the fraction of total annotations contained in analysis unit u.

As with Cohen's kappa, values of a >1 indicates agreement, 0 indicates the absence of agreement, and a <1 indicates systematic disagreement beyond what would be expected by chance.

There are many advantages to Krippendorff's alpha including the arbitrary definition of disagreement, the ability to account for missing data (e.g., the number of annotators can vary depending on the analysis unit), and its wide acceptance in the medical literature. Also, unlike Cohen's alpha, random chance disagreement, D_e, is determined empirically.

11.13 Alleviating the Cost of Gold Annotation

Because annotation is time-consuming and expensive, methods have been investigated to alleviate the cost of annotation, such as automatic pre-annotation and active learning. Pre-annotation involves automatically labeling the texts to be given to the human annotators, using an existing NLP system. The hope is that it will speed up the annotation process by providing preliminary annotations that will be corrected by the annotators. Two potential issues with this method are that it could introduce a bias in the annotation (e.g., annotators might over-rely on pre-annotations and will then miss the same kind of examples the automatic system misses) and that it could actually slow down the process, if the pre-annotation is of low quality. Ogren et al. (2008) used the MetaMap tool to pre-annotate clinical notes for disorders and observed that the annotation was slowed down because too much noise was generated by the tool.

Active learning is a machine learning technique that allows reducing the amount of necessary manual annotations, by selecting only the most "informative" instances to be annotated rather than sampling them randomly (Settles 2010). It typically uses only a small set of manually labeled examples to start with and a large set of unlabeled examples. The algorithm starts learning from the set of labeled examples, selects one or more informative instances (queries) from the unlabeled data to be labeled by a human annotator and then adds the newly labeled instance(s) to the training set, and repeats. Chen et al.(2012) successfully used active learning for annotating and classifying assertion of concepts in clinical notes. Miller et al. (2012) applied active learning to the clinical coreference task. Active learning also has potential drawbacks – for example, instances that the classifier finds "informative" tend to be more difficult for human annotators, and so even if active learning decreases the number of human-labeled instances required, per-instance labeling time may increase significantly, depending on the difficulty of the task, and so monitoring annotator time may be useful to ensure benefits of active learning.

D_e is calculated by averaging over all annotated pairs:
where $(i,u) \neq (i',u')$ and m are the number of annotators. Random chance disagreement is thereby defined by the average disagreement over all pairs regardless of their analysis unit or annotator.

11.14 NLP Components, Pipelines and Evaluation

An NLP system starts by receiving raw text as input, and has access to knowledge resources like the UMLS, statistical models it has built, and hand-specified rules encoding knowledge of the language in the domain. The output of a system may take a variety of forms, but there are several useful components that extract linguistic information that can be used in a variety of clinical applications. These components are often conceived as processing "pipelines," each operating on the input in turn and creating their own outputs. Core NLP components are shown in Table 11.1. They include tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference chains. Each step is described in Table 11.1.

Table 11.1 Core NLP components

Some core NLP tasks/ components	Description	Examples
Tokenization	A token is a group of characters that are grouped together to provide information. The process of dividing the input text into tokens	Sally became drowsy when she was taking carbmazepine
		/Sally//became//drowsy//when//she//was//taking//carbamazepine/
Sentence segmentation	Dividing written text in meaningful units.	//Sally became drowsy when she was taking carbmazepine////This occurs in 11.8 % of patients//
Part-of-speech tagging	Adding parts of speech to segmented text	NNP/Sally VBD/became JJ/drowsy IN/on NN/carbmazepine
		NNP = proper singular noun
		NN = Singular noun
		VBD = Verb, past tense
		JJ = Adjective
		IN = Preposition
Named entity extraction	Classifying text spans into predetermined categories like: diseases/disorders, signs/symptoms, anatomical sites, procedures, medications, people, temporal expressions, geographic locations	Sally was admitted to Cincinnati Children's Hospital Medical Center in Cincinnati OH on 4/20/2012
		Cincinnati Children's Hospital Medical Center
		Cincinnati OH
		4/20/2012

(continued)

Table 11.1 (continued)

Some core NLP tasks/ components	Description	Examples
Chunking	Sometimes called shallow parsing, usually representing sequential phrasal units.	Patient diagnosed with metastatic ascending colon cancer
		Noun phrase: metastatic ascending colon cancer; patient
		Verb phrase: diagnosed
Parsing	Analyzing text to determine grammatical structure.	nsubj(went-2, Sally-1) root(ROOT-0, went-2) det(hospital-5, the-4) prep_ to(went-2, hospital-5)
		partial output from Stanford parser. de Marneffe, 2008 (de Marneffe and Manning 2008). The Stanford Typed Dependencies Representation}
Coreference resolution	Resolving when multiple linguistic expressions refer to the same world entity.	Sally had severe side effects. She was admitted to the hospital.
		Sally had severe side effects. *She* was admitted to the hospital
Relation extraction	Finding semantic relations between named entities	Sally has been experiencing severe pain in the ball of her foot
		LocationOf(pain, ball of her foot)
		DegreeOf(pain, severe)
Temporal information extraction	Finding time ordering of events in patient history	Sally has been experiencing pain since August. Tumor was diagnosed in today's biopsy. She will begin chemotherapy after consultation next week
		BeginsOn(pain, August)
		Contains(today, diagnosed)
		Before(consultation, chemotherapy)
		Contains(next week, consultation)

There are several existing clinical NLP systems – MetaMap, YTex, TIES, OBO annotator, SemRep, Medlee, Apache clinical Text Analysis and Knowledge Extraction System (cTAKES (Aronson 2001; Friedman 2000; Garla and Brandt 2013; NLM(U.S) and NIH(U.S.) ; Savova et al. 2010; Source Forge ; UPMC)). MetaMap and OBO annotator focus on concept mapping and are tied to a specific ontology. SemRep extracts a set of six biomedically relevant relations from biomedical text. Medlee is a proprietary system that extracts entities and their associated

modifiers from clinical text; TIES is designed to operate on pathology notes and link them to tissue bank data. YTex does entity and select attribute recognition and has been integrated into Apache cTAKES. Apache cTAKES components perform duties as simple as locating sentence breaks and as complex as discovering relations between entities (e.g., the location of a tumor). cTAKES can also perform the extremely important task of identifying temporal events, dates and times – resulting in the absolute and relative placement of events in a patient timeline. cTAKES is currently being extended for deep phenotyping for the oncology domain (cancer.healthnlp.org) and is described in more detail in Chap. 12.

11.15 Evaluation

A number of pre-requisites are necessary for proper evaluation of NLP systems, including a strong gold standard and appropriate evaluation methods incorporating *performance measures* (Resnik and Lin 2010). Performance measures of an NLP system are generally geared toward the goals of a project, although the scope and quality of measures may vary within the context of a specific task (Jha 2011).

Several distinctions can be made in ways NLP systems are evaluated. First, *formative* evaluation should be distinguished from *summative* evaluation (Gale et al. 1992). Formative evaluation is performed (often iteratively) during the development of a system to measure progress, while summative evaluation assesses the performance of the system once it is complete. Second, *intrinsic* and *extrinsic* evaluations can be performed (Gale et al. 1992). Intrinsic evaluation assesses the quality of the output that the system is specifically designed to produce. Extrinsic evaluation measures the impact of the system on a task external to the system itself. Finally, another important distinction is made between *component-based* evaluation and *end-to-end* evaluation (Gale et al. 1992). As described above, NLP systems are often built from distinct components, and so performance can be measured focusing on each component individually (component evaluation) or on all components at once (end-to-end evaluation). Both evaluations are useful, as component evaluation allows researchers to understand the impact of each component, while end-to-end evaluation measures the effectiveness of the system in a real world setting. The performance of an automatic NLP system is usually assessed by comparing the system's output to a manually created reference gold standard. Section 11.10 described in detail the process of developing such a dataset. For evaluation purposes, it is split into training, development and test parts. The test part is the held-out data on which the final system is evaluated and results are reported. During the development phase, only the development split is utilized for tracking the performance.

An alternative to the training/development/test splits for statistical NLP systems is cross-validation (Cinchor 1992; Kohavi 1995). The classical type of cross-validation is k-fold cross validation. In k-fold cross validation, the annotated data is randomly partitioned into k subsets. K rounds of validation are then performed,

each time using a different subset as the validation data for testing the system and the remaining k−1 subsets for training the system. The k performance results can then be averaged to obtain a single performance measure. The advantage of this procedure is that it uses all data instances to validate a system while keeping the test sets independent (thereby preserving the integrity of the evaluation), and each instance is used only once for validation. In machine-learning-based NLP, 10-fold cross-validation is most commonly used, although leave-one-out cross-validation (i.e., N-fold cross-validation where N is the size of the data set) renders a less biased estimate of performance, particularly for small sample sizes (Braga-Neto and Dougherty 2004). It is further important that, if a sub-set of linguistic features are selected for analyses, that the selection of these features be based on the information contained within training set only (Hastie et al. 2005); selecting features based on, say, statistical tests with the entire data sample can lead to dramatic overestimations of performance. Even with such caution, cross-validation is more valuable as a *formative* evaluation than a *summative* evaluation, because during the development cycle repeated cross-evaluations amount to giving the researcher "peeks" at the test data. Therefore, it is important to set aside a test set from the gold standard data that is only used for a final experiment to obtain a more reliable estimate of performance on new data.

Standard metrics to evaluate the performance of (medical and general-language) NLP systems are Recall (R), Precision (P) and F1-measure (F). Recall (more commonly called *sensitivity* in medical literature), is the number of instances correctly predicted by a system divided by the total number of instances in the gold standard. Precision (called *positive predictive value* in medical literature) is the number of correctly predicted instances divided by the total number of instances predicted by a system. F-measure is the weighted harmonic mean of precision and recall (as detailed in Sect. 11.15).

Accuracy is defined as (Fig. 11.5)

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Fig. 11.5

where t_p, t_n, f_p, and f_n are the number of true positives, true negatives, false positive, and false negatives, respectively. The accuracy from random chance in a two class task always corresponds to a value of 0.5. Although widely used and intuitive, it is important to note for imbalanced data, a naïve classifier that only predicts a single class can produce a high accuracy.

In contrast, the F_β, as discussed in [PREVIOUS SECTION], gives no weight to true negatives, and so this situation is avoided at the cost of the random baseline value being dependent on the balance of the data set.

Even in cases where the classifier must ultimately produce a binary prediction, evaluating the classifier in term of its ability to rank objects can be useful. To demonstrate,

suppose a classifier is built that uses a patient's words to determine whether they are suffering from depression. Based on a patient's word frequencies, the classifier outputs a number between 0 and 1; numbers closer to 1 indicate a higher probability of depression. A threshold, dependent on the desired precision and recall, is then set on the output. All patients above that threshold are then classified as "depressed", and those below the threshold are classified as "not depressed". However, the decision about whether the classifier should alert the clinician to the patient's depression may depend on the clinical application and even the clinician. It is then important to give a more general measure of how well the classifier behaves without committing to a certain precision/recall.

The area under the receiver operating curve (AROC) renders such a general measure (Hanley and McNeil 1982; Lasko et al. 2005). The receiver operating curve (ROC) is generated by evaluating the true positive and false positive rates from various classifier thresholds. The true positive rate is then plotted against the false positive rate for each threshold, and the area under the resulting curve (AROC) is evaluated. The AROC is then the probability that a randomly selected data point from the positive category (e.g., "depressed" patients) will have a classifier output value larger than a randomly selected data point from the negative category (e.g., "not depressed" patients). An AROC of 0.5 corresponds to the expected performance of a randomly guessing classifier.

While the AROC does not directly quantify the performance of a specific set of binary classifications, it does quantify the overall performance of the classifier. Also, it can directly quantify the performance of the classifier in tasks where the prediction is a rank. Further, the AROC does not need to be corrected for imbalanced data sets (the AROC is always 0.5 for purely random classifier outputs), and there exist analytic methods of calculating the standard error (Hanley and McNeil 1982) and comparing pairs of AROCs (DeLong et al. 1988).

11.16 Shared Tasks in Clinical NLP

There have been several efforts to provide benchmarks to evaluate NLP systems. The Text Retrieval Conference (TREC) (NIST), the Text Analysis Conferences (TAC) (Dy and Brodley 2004), the Computational Natural Language Learning Conferences (ACL and SIGNLL 2010) and Senseval (CDC 2007; SENSEVAL) have provided independent competitive evaluation of many of the core NLP components.

Competitive shared tasks were designed to encourage development of medical NLP systems and applications and to evaluate them in a standard framework. The 2007 NLP challenge of the Computational Medicine Center (CMC (2007)) was the first shared task in the medical domain and focused on automatic mapping of ICD-9-CM codes to radiology free-text reports (Pestian et al. 2007). The shared tasks organized by i2b2 (i2b2 2012) spanned seven years and provided annual bakeoffs on a variety of tasks: on de-identification of clinical text (Uzuner et al. 2007), on

identification of patient smoking status (Uzuner et al. 2008), on recognition of obesity and co-morbidities (Uzuner 2009), on medication information extraction (Uzuner et al. 2010), on identification of concepts, assertions and relations (Uzuner et al. 2011), on coreference resolution (Uzuner et al. 2012) and temporal relations (Sun et al. 2013). The 2011 i2b2 challenge was held in conjunction with the Computational Medicine Center which proposed a second shared task on classification of emotions in suicide notes (Pestian et al. 2012). The 2011 TREC National Institute for Standards in Technology (NIST), which organizes annual evaluations on a variety of information retrieval tasks, had a Medical Records Track focused on the retrieval of patients relevant to specific criteria (i.e. specific diseases and treatments).

Recent developments have been the introduction of clinical NLP tasks to the general computer science community. There are four shared tasks/challenges, each introducing a new facet of the ShARe annotated corpus: Conference and Labs of the Evaluation Forum/Shared Annotated Resources (CLEF/ShAREe 2013) (CLEF/ShAREe 2014); SemEval 2014 Analysis of Clinical Text Task 7 (ALT Server – QCRI 2014; Pradhan et al. 2014); and SemEval 2015 Analysis of Clinical Text Task 14 (ALT Server – QCRI 2015). The SemEval 2014 and 2015 tasks were in the top three most popular tasks in terms of number of participants (from over 20 SemEval shared tasks) demonstrating the growing interest in analysis of clinical text.

Further pushing the boundary of developing methods for temporal relation extraction from the clinical text are the 2015 and 2016 Clinical SemEval tasks (ALT Server – QCRI 2014; 2015; Bethard et al. 2015). Clinical TempEval brings these temporal information extraction tasks to the clinical domain, using clinical notes and pathology reports from the Mayo Clinic. This follows recent interest in temporal information extraction for the clinical domain, e.g., the i2b2 2012 shared task (Sun et al., 2013), and broadens our understanding of the language of time beyond the domain of newswire text.

11.17 Challenges in Clinical NLP

Despite previous successes in applying NLP to extract information from unstructured clinical narratives, a number of research questions are still open for exploration and they mainly lie in higher level discourse tasks requiring not only the processing of the text but its understanding. Although tremendous efforts have been made to ensure its accuracy, the clinical description is inevitably ambiguous and diverse due to semantic (e.g., homonyms, synonyms, and acronyms) and syntactic (e.g., polarity, temporality, and certainty) characteristics of human language. Disambiguation of clinical text requires sophisticated methods just like any other higher-level discourse task. Gold annotations provide the much needed training and evaluation data. The creation of such datasets is labor-intensive (Zweigenbaum et al. 2007) and sharing such corpora is often infeasible due to patient privacy and confidentiality requirements (Sect. 11.4). In addition, the literature research has

shown that data-driven NLP systems trained on corpora from one institution will only be partially transferable to other institutions (Deleger et al. 2014). Domain adaptation with minimal supervision has become a new actively researched area enabled by the availability of big data. Although portability of NLP methods across domains is currently limited by the data the methods are trained on, the minimally supervised technologies hold enormous promise for providing effective and efficient scalable solution in the very near future.

References

AAMT. American Association for Medical Transcription position paper. Quality assurance guidelines. J Am Assoc Med Transcr. 1994;13(6):33–7.

AAMT. American Association for Medical Transcriptionists. Best practices for measuring quality in medical transcription March 2015. 2005. Retrieved from http://www.startranscriptions.com/QualityMeasurementMT.pdf.

ACL and SIGNLL. Association of Computational Linguistics (ACL) and Special Interest Group on Natural Language Learning (SIGNLL). CoNLL: the conference of SIGNLL; 2010. Retrieved from http://ifarm.nl/signll/conll/.

Albright D, Lanfranchi A, Fredriksen A, Styler WF, Warner C, Hwang JD, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. J Am Med Inform Assoc. 2013 Sep 1;20(5):922–30.

ALT Server – QCRI. SemEval – 2014 Task 7. 2014. Retrieved from http://alt.qcri.org/semeval2014/task7/.

ALT Server – QCRI. SemEval-2015 Task 14: analysis of clinical text. 2015. Retrieved from http://alt.qcri.org/semeval2015/task14/.

Altman D. Inter-rater agreement. Practical statistics for medical research. 1991;5:403–409.

Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Paper presented at the proceedings of the AMIA symposium. 2001.

Aston G, Burnard L. The BNC handbook: exploring the British National Corpus with SARA: Capstone. 1998.

Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet project. Poster presented at the proceedings of the 17th international conference on computational linguistics – Volume 1, Montreal; 1998. http://acl.ldc.upenn.edu/C/C98/C98-1013.pdf.

Banerjee M, Capozzoli M, McSweeney L, Sinha D. Beyond kappa: a review of interrater agreement measures. Can J Stat. 1999;27(1):3–23.

Bethard S, Derczynski L, Savova G, Savova G, Pustejovsky J, Verhagen M. Semeval-2015 task 6: Clinical tempeval. Proc SemEval. 2015.

Birdwhistell RL. Kinesics and context: essays on body motion communication. Philadelphia: University of Pennsylvania press; 2010.

Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. J Biomed Inform. 2003;36(6):414–32.

Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics. 2004;20(3):374–80.

brat. brat rapid annotation tool. Retrieved from http://brat.nlplab.org/.

Bright W. International encyclopedia of linguistics. New York: Oxford University Press; 1992.

Brownstein JS, Freifeld CC, Madoff LC. Influenza A (H1N1) virus, 2009 – online monitoring. N Engl J Med. 2009;360(21):2156.

Brownstein JS, Sordo M, Kohane IS, Mandl KD. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. PLoS One. 2007;2(9):e840.

cancer.healthnlp.org. Health NLP. Retrieved from https://healthnlp.hms.harvard.edu/cancer/wiki/index.php/Main_Page.
CDC. Suicide trends among youths and young adults aged 10–24 years – United States, 1990–2004. MMWR Morb Mortal Wkly Rep. 2007;56(35):905–8.
Chapman WW, Dowling JN. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. J Biomed Inform. 2006;39(2):196–208.
Chapman WW, Dowling JN, Hripcsak G. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. Int J Med Inform. 2008;77(2):107–13.
Chen W-T, Styler W. Anafora: a web-based general purpose annotation tool. Paper presented at the Proceedings of the North American Association for Computational Linguistics conference, Atlanta; 2013.
Chen Y, Mani S, Xu H. Applying active learning to assertion classification of concepts in clinical text. J Biomed Inform. 2012;45(2):265–72.
Chomsky N. Aspects of the theory of syntax. Cambridge: M.I.T. Press; 1965.
Cieri C, Miller D, Walker K. The fisher corpus: a resource for the next generations of speech-to-text. Paper presented at the LREC, vol. 4. 2004; p. 69–71.
Cinchor N. The statistical significance of MUC4 results. Paper presented at the MUC4 '92 proceedings of the 4th conference on message understanding. 1992.
CLAN. Child language data exchange system. Retrieved February 12, 2016 http://childes.psy.cmu.edu/Clan/.
CLEF/ShAREe. Sharing annotated resources. 2013. Retrieved from https://sites.google.com/site/shareclefehealth/
CLEF/ShAREe. CLEF ehealth 2014: lab overview. 2014. Retrieved from http://clefehealth2014.dcu.ie/.
CMC. Computational medical center. 2007 international challenge: classifying clinical free text using natural language processing. 2007. Retrieved from http://computationalmedicine.org/challenge/previous.
Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
Deleger L, Lingren T, Ni Y, Kaiser M, Stoutenborough L, Marsolo K, Kouril M, Molnar K, Solti I. Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. J Biomed Inform. 2014;50:173–83.
DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42(5):760–72.
Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010;26(9):1205–10.
Desmet, B. Finding the online cry for help: automatic text classification for suicide prevention. PhD Dissertation. Ghent: Ghent University; 2014.
Dy JG, Brodley CE. Feature selection for unsupervised learning. J Mach Learn Res. 2004;5:845–89.
Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. Answering physicians' clinical questions: obstacles and potential solutions. J Am Med Inform Assoc. 2005;12(2):217–24.
eMERGE Network. A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. Retrieved from http://gwas.net/.
Fellbaum C, Grabowski J, Landes S. Performance and Confidence in a Semantic Annotation Task. In: Fellbaum C, editor. WordNet: an electronic lexical database. Cambridge, MA: MIT Press; 1998.
Friedman C. A broad-coverage natural language processing system. Paper presented at the proceedings of the AMIA symposium. 2000.

Gale W, Church KW, Yarowsky D. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. Poster presented at the proceedings of the 30th annual meeting on association for computational linguistics, Newark, Delaware. 1992.

Garla VN, Brandt C. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. J Am Med Inform Assoc. 2013;20(5):882–6.

Geisser S. Predictive sample reuse method with applications. J Am Stat Assoc. 1975;70(350):320–8.

Giacomini KM, Brett CM, Altman RB, Benowitz NL, Dolan ME, Flockhart DA, Johnson JA, Hayes DF, Klein T, Krauss RM, Kroetz DL, McLeod HL, Nguyen AT, Ratain MJ, Relling MV, Reus V, Roden DM, Schaefer CA, Shuldiner AR, Skaar T, Tantisira K, Tyndale RF, Wang L, Weinshilboum RM, Weiss ST, Zineh I. The pharmacogenetics research network: from SNP discovery to clinical drug response. Clin Pharmacol Ther. 2007;81(3):328–45.

Gomez JM. Language technologies for suicide prevention in social media. Paper presented at the 5th information systems research working days (JISIC 2014). 2014.

Grouin C, Rosset S, Zweigenbaum P, Fort K, Galibert O, Quintard L. Proposal for an extension of traditional named entities: from guidelines to evaluation, an overview. Poster presented at the proceedings of the 5th linguistic annotation workshop (LAW V '11), Portland, Oregon. 2011.

Gwet KL. On Krippendorff's Alpha coefficient. Advanced analytics LLC inter-rater reliability Publicaitons. 2011. Retrieved from http://www.agreestat.com/research_papers/onkrippendorf-falpha_rev10052015.pdf.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.

Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. Math Intell. 2005;27(2):83–5.

Health Map. 2006. Retrieved from http://www.healthmap.org/en/.

Hicks J. The potential of claims data to support the measurement of health care quality Santa Monica. Santa Monica: RAND Corporation; 2003. Retrieved from http://www.rand.org/pubs/rgs_dissertations/RGSD171.

Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005;12(3):296–8.

Huang Y-P, Goh T, Liew CL. Hunting suicide notes in web 2.0-preliminary findings. Paper presented at the multimedia workshops, 2007 ISMW'07 Ninth IEEE international symposium on. 2007.

i2b2. Informatics for integrating biology & the bedside. Datasets. Retrieved from https://www.i2b2.org/NLP/DataSets/Main.php.

i2b2. Informatics for integrating biology & the bedside. 2012 NLP shared task: shared-tasks and workshop on challenges in natural language processing for clinical data. 2012. Retrieved from https://www.i2b2.org/NLP/HeartDisease/.

Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, Argyle T. Tracking suicide risk factors through Twitter in the US. Crisis. 2014;35(1):51–9.

Jha AK. The promise of electronic records: around the corner or down the road? JAMA. 2011;306(8):880–1.

Jones K. Natural language processing: a historical review [Paper]. Current issues in computational linguistics: in honour of Don Walker. 2001. Retrieved from http://www.cl.cam.ac.uk/archive/ksj21/histdw4.pdf.

Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River: Prentice Hall; 2000.

Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, Crane PK, Pathak J, Chute CG, Bielinski SJ, Kullo IJ, Li R, Manolio TA, Chisholm RL, Denny JC. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med. 2011;3(79):79re71.

Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12(6):417–28.

Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Poster presented at the proceedings of the fourteenth international joint conference on artificial intelligence. 1995.

Krippendorff K. Computing Krippendorff's alpha reliability. Departmental papers (ASC). 2007;43.

Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform. 2005;38(5):404–15.

Li Q, Deleger L, Lingren T, Zhai H, Kaiser M, Stoutenborough L, Jegga AG, Cohen KB, Solti I. Mining FDA drug labels for medical conditions. BMC Med Inform Decis Mak. 2013a;13(1):1.

Li T, Ng B, Chau M, Wong P, Yip P. Collective intelligence for suicide surveillance in web forums intelligence and security informatics. Berlin: Springer; 2013b. p. 29–37.

Luo Z, Johnson SB, Lai AM, Weng C. Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. AMIA Annu Symp Proc. 2011;2011:843–52.

MacWhinney B. The CHILDES project: tools for analyzing talk, The database, vol. 2. 3rd ed. Mahwah: Lawrence Erlbaum Associates; 2000.

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: the next frontier for innovation, competition, and productivity. Washington, DC: McKinsey Global Institute; 2011.

Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: the Penn Treebank. Comput Linguist. 1993;19(2):313–30.

Matykiewicz P, Duch W, Pestian J. Clustering semantic spaces of suicide notes and newsgroups articles. Poster presented at the proceedings of the workshop on current trends in biomedical natural language processing, Boulder; 2009.

McCarty C, Chisholm R, Chute C, Kullo I, Jarvik G, Larson E, Li R, Masys D, Ritchie M, Roden D, Struewing J, Wolf W, Team, t. e. The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;4(1):13.

Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005;12(4):448–57.

Meyers A, Reeves R, Macleod C, Szekely R, Zielinska V, Young B, Grishman R. The NomBank project: an interim report. Paper presented at the HLT-NAACL 2004 workshop: Frontiers in Corpus Annotation. 2004.

Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128–144.

Miller T, Dligach D, Savova GK. Active learning for coreference resolution. Poster presented at the BioNLP workshop at the conference of the North American Association of Computational Linguistics (NACCL), Montreal. 2012.

Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, Dittus RS, Rosen AK, Elkin PL, Brown SH, Speroff T. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA. 2011;306(8):848–55.

National Institutes of Health and National Institute of General Medical Sciences. The NIH Pharmacogenomics Research Network (PGRN). Retrieved from http://www.nigms.nih.gov/Research/FeaturedPrograms/PGRN/.

National Research Council. "Recommendations." Language and machines: computers in translation and linguistics. 1966. Retrieved from http://www.nap.edu/openbook.php?isbn=ARC000005.

NIST. National Institute for Standards in Technology. Text REtrieval Conference (TREC). Retrieved from http://trec.nist.gov/.

NLM(U.S) and NIH(U.S.). SemRep: semantic knowledge representation. Retrieved from http://semrep.nlm.nih.gov/.

Ogren, P. Knowtator: a protégé plug-in for annotated corpus construction. Paper presented at the proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. 2006.

Ogren PV, Savova GK, Chute C. Constructing evaluation corpora for automated clinical named entity recognition. Paper presented at the proceedings of the sixth international conference on language resources and evaluation (LREC '08), Marrakech, Morocco. 2008.
Osborne JW. Best practices in quantitative methods. Thousand Oaks: Sage Publications; 2008.
Owen JE, Giese-Davis J, Cordova M, Kronenwetter C, Golant M, Spiegel D. Self-report and linguistic indicators of emotional expression in narratives as predictors of adjustment to cancer. J Behav Med. 2006;29(4):335–45.
Pakhomov SV, Coden A, Chute CG. Developing a corpus of clinical notes manually annotated for part-of-speech. Int J Med Inform. 2006;75(6):418–29.
Palmer M, Dang HT, Fellbaum C. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. Nat Lang Eng. 2007;13(02):137–63.
Palmer M, Gildea D, Kingsbury P. The proposition bank: an annotated corpus of semantic roles. Comput Linguist. 2005;31(1):71–106.
Palmer M, Xue N. Linguistic annotation. In: Clark A, Fox C, Lappin S, editors. The handbook of computational linguistics and natural language processing. Chichester/Malden: Wiley-Blackwell; 2010. p. 13–21.
Penn Treebank Project. Retrieved from http://www.cis.upenn.edu/~treebank/.
Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001. Mahwah: Erlbaum Publishers; 2001 (www.erlbaum.com).
Pennebaker JW, Mayne TJ, Francis ME. Linguistic predictors of adaptive bereavement. J Pers Soc Psychol. 1997;72(4):863.
Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, Duch W. A shared task involving multi-label classification of clinical free text. Poster presented at the proceedings of the workshop on BioNLP 2007: biological, translational, and clinical language processing, Prague, Czech Republic. 2007.
Pestian JP, Matykiewicz P, Grupp-Phelan J, Lavanier SA, Combs J, Kowatch R. Using natural language processing to classify suicide notes. AMIA Annu Symp Proc. 2008:1091.
Pestian JP, Matykiewicz P, Linn-Gust M, South B, Uzuner O, Wiebe J, Cohen KB, Hurdle J, Brew C. Sentiment analysis of suicide notes: a shared task. Biomed Inform Insights. 2012;5 Suppl 1:3–16.
Pestian J P, Sorter M, Cohen KB, McCullumsmith C, Gee JT, Morency LP, Scherer S, Rohlfs LftSRG. A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial [in press]. Suicide Life Threat Behav. 2016.
Poesio M. Discourse annotation and semantic annotation in the GNOME corpus. Poster presented at the proceedings of the 2004 ACL workshop on discourse annotation, Barcelona, Spain. 2004. http://delivery.acm.org/10.1145/1610000/1608948/p72-poesio.pdf?ip=205.142.197.101&acc=OPEN&CFID=102282657&CFTOKEN=85566326&__acm__=1336683135_1b135c40faae93e87b9f6cbfbfe6d031.
Poesio M, Vieira R. A corpus-based investigation of definite description use. Comput Linguist. 1998;24(2):183–216.
Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G. Semeval-2014 task 7: analysis of clinical text. SemEval. 2014;199(99):54.
Prasad R, Dinesh N, Lee A, Miltsakaki E, Robaldo L, Joshi AK, Webber BL. The Penn Discourse TreeBank 2.0. Paper presented at the proceedings of LREC (Language Resources and Evaluation Conference). 2008.
PropBank Project. Retrieved from http://verbs.colorado.edu/~mpalmer/projects/ace.html.
Pustejovsky J, Hanks P, Sauri R, See A, Gaizauskas R, Setzer A, Radev D, Sundheim B, Day D, Ferro L, Lazo M. The TIMEBANK corpus. Paper presented at the proceedings of the corpus linguistics. 2003.
Pustejovsky J, Stubbs A. Natural language annotation for machine learning. Sebastopol: O'Reilly Media; 2012.

Ratner NB, Rooney B, MacWhinney B. Analysis of stuttering using CHILDES and CLAN. Clinical Linguistics & Phonetics. 1996;10(3):169–87.
Resnik P, Lin J. Evaluation of NLP systems. In: Clark A, Fox C, Lappin S, editors. The handbook of computational linguistics and natural language processing. Chichester/Malden: Wiley-Blackwell; 2010. p. 271–96.
Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, Setzer A. Building a semantically annotated corpus of clinical texts. J Biomed Inform. 2009;42(5):950–66.
Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–13.
SENSEVAL. Evaluation Exercises for the Semantic Analysis of Text. Retrieved from http://www.senseval.org/.
Settles B. Active learning literature survey. Madison: University of Wisconsin; 2010;52(55–66), 11.
Shannon CE, Weaver W. The mathematical theory of communication. Urbana: University of Illinois Press; 1949.
ShaRe. Shared Annotated Resource for the Clinical Domain. Clinical NLP Annotation Retrieved from https://www.clinicalnlpannotation.org/index.php/Main_Page.
SHARPn.org. Strategic Health IT Advanced Research Projects (SHARP): research focus area 4. Retrieved from http://informatics.mayo.edu/sharp/index.php/Main_Page.
Silverman KE, Beckman ME, Pitrelli JF, Ostendorf M, Wightman CW, Price P, Pierrehumbert JB, Hirschberg J. TOBI: a standard for labeling English prosody. Paper presented at the proceedings of ICSLP. 1992.
Solti I, Cooke CR, Xia F, Wurfel MM. Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches. Proceedings (IEEE Int Conf Bioinformatics Biomed). 2009, p. 314–319.
Source Forge. OBO Annotator. Retrieved from https://sourceforge.net/projects/obo-annotator/.
South BR, Shen S, Jones M, Garvin J, Samore MH, Chapman WW, Gundlapalli AV. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinform. 2009;10(Suppl 9):S12.
Styler IV WF, Bethard S, Finan S, Palmer M, Pradhan S, de Groen PC, Erickson B, Miller T, Lin C, Savova G. Temporal annotation in the clinical domain. Transactions of the Association for Computational Linguistics. 2014;2:143–54.
Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. J Am Med Inform Assoc. 2013;20(5).
Thompson P, Poulin C, Bryan CJ. Predicting military and veteran suicide risk: cultural aspects. ACL. 2014;2014:1.
THYME. Temporal History of Your Medical Events. Retrieved from http://clear.colorado.edu/compsem/index.php?page=endendsystems&sub=temporal.
TimeML. TimeML specifications. Retrieved from http://www.timeml.org/publications/specs.html.
TLA. The Language Archive. ELAN: a professional tool for the creation of complex annotations on video and audio resources. Retrieved February 12, 2016 https://tla.mpi.nl/tools/tla-tools/elan/.
Tomanek K, Hahn U. Timed annotations – enhancing MUC7 metadata by the time it takes to annotate named entities. Paper presented at the proceedings of the linguistic annotation workshop. 2009.
UMLS [Internet]. United Medical Language System (UMLS). Retrieved from https://www.nlm.nih.gov/research/umls/.
UPMC. University of Pittsburg Medical Center. TIES: a clinical text search engine. Retrieved from http://ties.upmc.com/.
Uzuner O. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc. 2009;16(4):561–70.

Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc. 2012;19(5):786–91.
Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc. 2007;14(5):550–63.
Uzuner O, Sibanda TC, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. Artif Intell Med. 2008;42(1):13–35.
Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. J Am Med Inform Assoc. 2010;17(5):519–23.
Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011;18(5):552–6.
Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. Lang Resour Eval. 2005;39(2):165–210.
Wittenburg P, Brugman H, Russel A, Klassmann A, Sloetjes H. Elan: a professional framework for multimodality research. Paper presented at the proceedings of LREC. 2006.
WordNet. A lexical database for English. Retrieved June 1, 2012, from Princeton University http://wordnet.princeton.edu/.
Yetisgen-Yildiz M, Solti I, Xia F. Using amazon's mechanical turk for annotating medical named entities. AMIA Annu Symp Proc. 2010;2010:1316.
Young EC. The effects of treatment on consonant cluster and weak syllable reduction processes in misarticulating children. Language, Speech, and Hearing Services in Schools. 1987;18(1):23–33.
Zampolli A, Calzolari N, Palmer MS, Walker DE. Current issues in computational linguistics : in honour of Don Walker. Pisa/Norwell: Giardini ; Distributed in the U.S.A. and Canada by Kluwer Academic Publishers; 1994.
Zhan C, Miller MR. Administrative data based patient safety research: a critical review. Qual Saf Health Care. 2003:12(Suppl 2):ii58–63.
Zhang L, Huang X, Liu T, Li A, Chen Z, Zhu T. Using linguistic features to estimate suicide probability of Chinese microblog users human centered computing. New York: Springer; 2014. p. 549–59.
Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Brief Bioinform. 2007;8(5):358–75.

Chapter 12
Natural Language Processing: Applications in Pediatric Research

Guergana Savova, John Pestian, Brian Connolly, Timothy Miller, Yizhao Ni, and Judith W. Dexheimer

Abstract We discuss specific biomedical Natural Language Processing-based applications that cover a wide spectrum of use cases within the field of translational and health services research. In our uses cases we focus on four categories of applications: (1) Information Extraction (IE), (2) Document Classification, (3) Patient Classification, and (4) Sentiment Analysis. We show how the extracted information could be used for (a) Phenotype identification, (b) Comparative effectiveness studies, (c) Cohort identification, (d) Meaningful Use, and (e) Linking patients' phenotype and genotype. In addition, we discuss the use of Natural Language Processing components for de-identification of large collections of patient notes. We review the

G. Savova, Ph.D. (✉)
Children's Hospital Boston and Harvard Medical School,
300 Longwood Avenue, Enders 138, Boston, MA 02115, USA
e-mail: Guergana.Savova@childrens.harvard.edu

J. Pestian, Ph.D., M.B.A. (✉)
Department of Pediatrics and Biomedical Informatics, Division of Emergency Medicine,
Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine,
3333 Burnet Ave, ML-2008, Cincinnati, OH 45229-3039, USA

B. Connolly, Ph.D.
Department of Pediatrics, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Avenue, Cincinnati, OH 45229-3039, USA

T. Miller, Ph.D.
Department of Pediatrics, Harvard Medical School, Boston Children's Hospital,
300 Longwood Avenue Enders 138, Boston, MA 02115, USA

Y. Ni, Ph.D.
Department of Pediatrics and Biomedical Informatics, Division of Biomedical Informatics,
Children's Hospital Medical Center, University of Cincinnati College of Medicine,
3333 Burnet Avenue, ML-7024, Cincinnati, OH 45229-3039, USA

J.W. Dexheimer, Ph.D.
Departments of Pediatrics and Biomedical Informatics, Divisions of Emergency Medicine and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, 3333 Burnet Ave, ML-2008, Cincinnati, OH 45229, USA

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_12

literature for examples of pediatric natural language processing applications and show the transferability of select adult clinical natural language processing applications to the pediatric population.

Keywords Classification • Cohort • De-identification • Information extraction • Natural language processing • NLP • Phenotype • Protected health information • PHI

12.1 Overview

As Chap. 9 on the basic principles of Natural Language Processing (NLP) pointed out, NLP focuses on developing technologies for our most ubiquitous product: human language (Coursera.org 2012). The electronic version of that product has exploded because of pervasive computing, and interest in and applications of NLP have been unprecedented. In the fall of 2011, Stanford University offered the first free world-wide online class in NLP taught by renowned NLP experts Profs. Jurafsky and Manning. The reaction of the student community has been nothing short of amazing – the class drew an enrollment in the tens of thousands of students, unofficial numbers are 60,000 students (Online Colleges.net. 2012).

In this section we outline some of the technologies enabling clinical translational and health care research. Two of the most well-known applications of NLP technologies are Information Retrieval (IR) and Question Answering (QA). IR systems focus on developing methods and algorithms for retrieving documents from a given repository (Jurafsky and Martin 2000; Manning et al. 2000). IR has become mainstream through search engines like Google, Bing and Yahoo!. An IR engine returns a set of documents ranked by an algorithm, the end user is then expected to sift through them to find the relevant information. On the other hand, QA systems aim to present results at a finer level, usually at the level of the sentence or the paragraph.

In 2011, IBM introduced to the world their Watson QA system and demonstrated that it can outsmart a highly knowledgeable human by competing in, and winning, the quiz show *Jeopardy!* (IBM 2012). The Watson QA system showcased the end product of a well-thought out and efficient software architecture weaving NLP technologies (constituency parsers, dependency parsers, entity recognizers, relation extractors, co-reference modules, semantic role labelers, expected answer classifiers), lexical resources (a variety of encyclopedias and gold standard parsed data off which machine learners were built) and purely engineering solutions (Apache Unstructured Information Management Architecture (UIMA 2012)) and fast computing hardware/databases.

Clinical QA systems are specifically tailored to the domain of medicine (for an overview of the QA for biomedicine consult (Athenikos and Han 2010)). Demner-Fushman and colleagues (2007) developed a system of feature extractors that identify subject populations, interventions, and outcomes in medical studies. Later they

extended this research to create a general-purpose QA system. Weiming et al. (2007) experimented with the Unified Medical Language System (UMLS) (Lindberg et al. 1993) semantic relationships and created a system that extracts exact answers from the output of a Lucene-based (Lucene 2012) IR system; however, their approach requires an exact semantic match, which limits performance on more complex and non-factoid questions. Athenikos et al. (2009) propose a rule-based system based on human-developed question/answer UMLS semantic patterns; such an approach is likely to perform well on specific question types but might have difficulty generalizing to new questions. AskHermes (Yu and Cao 2008) finds answers based on keyword searching. The Multi-source Integrated Platform for Answering Clinical Questions (MIPACQ) also uses the UMLS as the backbone ontology and returns answers at the paragraph level (Cairns et al. 2011; Nielsen and et al. 2010).

A more middleware application is Information Extraction (IE). IE converts the seemingly unstructured natural language (although linguistic studies show that language per se has inherent expressible structure) into explicitly structured computable data elements. For example, the text in the following sentence "Patient was diagnosed with cancer located in the ascending colon" will be converted to a computable event data structure anchored at the event of *cancer,* where the experiencer is the *patient* and the anatomic location is the *ascending colon*. This data structure is then used as the text metadata tags which are stored in databases to facilitate retrieval and data mining.

A pioneering clinical IE system is the Medical Language Extraction and Encoding System (MedLEE) (Friedman 1997, 2000; Hripcsak et al. 1998) developed at Columbia University. MedLEE has been applied to many use cases, however it is not open-sourced. Other examples of IE systems (focusing on English clinical text) are University of Sidney's Health Information Technologies Research Laboratory's NLP tool suite (Health Information Technologies Research Laboratory 2012), Health Information Text Extraction (2012; Zeng et al. 2006), IBM's MedTAS (Mack et al. 2004), SymText and MPLUS (Christensen et al. 2002; Haug and et al. 1995), and University of Pittsburgh's Tissue Information Extraction System (TIES) (cancer Text Information Extraction System 2012; Crowley et al. 2010), NOBLE coder (Tseytlin et al. 2016). Other tools developed primarily for processing biomedical scholarly articles include the National Library of Medicine MetaMap (Aronson and et al. 2000), providing mappings to the UMLS Metathesaurus concepts, those from the National Center for Text Mining (2012), JULIE lab (2012), and U-compare (2012), with some applications to the clinical domain (Meystre and Haug 2005). Tseytlin et al. (2016) offer a comparison across multiple systems.

A mature NLP IE system is the Apache top-level project Clinical Text Analysis and Knowledge Extraction System (cTAKES) platform (2012; Savova et al. 2010a). cTAKES is built within the UIMA engineering framework, which provides a solid basis for scalability, expandability and collaborative software development. The overarching principle is *process once, use many times* thus allowing pre-processing of the data, the result of which is the generation of indices to be used for retrieval purposes spanning a variety of use cases – from phenotype discovery, cohort identification, patient care to "meaningful use" of the EHR and comparative effectiveness. These topics are discussed in the subsequent subsections.

The machine learning components of cTAKES are trained on gold annotations of clinical text. cTAKES processes clinical narratives and identifies clinical named entity mentions (NEs) of types Drugs, Diseases/Disorders, Signs/Symptoms, Anatomical sites, and Procedures per UMLS semantic types and groups. Each discovered NE is assigned attributes such as text span, ontology mapping code (SNOMED CT, RxNORM), context (family history of, current, unrelated to patient), and a negation indicator. The core cTAKES components include following annotators in this specific order (for details, consult Savova et al. (2010a)):

- Sentence boundary detector, which is a wrapper around OpenNLP's sentence boundary detector (OpenNLP Tools 2012), but trained on clinical data. The module finds the end of the sentence.
- Rule-based tokenizer to separate punctuations from words.
- Normalizer, which is wrapper around LVG (Lexical Systems Group 2012) to standardize for example morphologically different phrases with same meaning, e.g. "infection", "infecting", "infects" normalize to the same form of "infect".
- Context dependent tokenizer grouping tokens to create Date, Digits and other types of groupings.
- Part-of-speech tagger, which is a wrapper around OpenNLP's but trained on clinical data. The module assigns a part-of-speech label to each word, e.g. "infects" is a verb.
- Phrasal chunker, which is a wrapper around OpenNLP's but trained on clinical data to detect phrases such as noun phrases, verb phrases, prepositional phrases. Most concepts are expressed as noun phrases thus the noun phrases become the lookup window for the next module.
- Dictionary lookup annotator that performs a permutational lookup against a dictionary database. The permutations are computed off the components within the lookup window, which is the noun phrase.
- Context annotator based on NegEx (Chapman et al. 2001) to link information with the patient or family members.
- Two negation discovery modules for the user to choose from: (1) Negation detector based on NegEx to discover whether a concept is negated or not. (2) a machine learning module (Wu et al. 2014).
- Dependency parser (Choi and Palmer 2011a, b) to detect dependency relations between words.
- Semantic role labeler based on the dependency parse labels to parse the sentence into a predicate with its arguments (Choi and Palmer 2011b). For example, "In 2007, patient was diagnosed with autism", there is a predicate associated with "diagnosed" which has three arguments: "the patient", "autism" and "2007" each associated with a semantic role (Palmer et al. 2005).
- Drug mention annotator which is a module to populate the drug template with attributes for change status (started, discontinued, increased, decreased, no change), dosage, duration, end date, form, frequency, route, start date, strength.
- Temporality module to extract temporal events, expressions and the relations among them to create a patient's timeline (Lin et al. 2016).

There is an additional module for the identification of the patient smoking status (Sohn et al. 2010): past smoker, current smoker, non-smoker, smoker, unknown. In

a typical IE application scenario, the SNOMED CT and RxNORM codes generated by cTAKES are retained as indices to the clinical narrative. Any subsequent queries directly go against these indices, obviating real time processing of the text. An NLP-based IE application thus creates metadata tags used for indexing and retrieval.

The open source and modular nature of cTAKES facilitates the possibility that each of the NLP components within cTAKES can be re-purposed within the architecture of a variety of applications such as QA and patient summarization to name a few. Such repurposing is already under way to port cTAKES to other medical domains (e.g. extracting information from clinical trial announcement text (Deleger et al. 2012)). cTAKES is also being extended for deep phenotyping for the oncology domain (cancer.healthnlp.org. 2016). It has been used in a number of translational research project which are described in more detail below.

12.2 NLP-Based Automated De-Identification of Clinical Narrative Text

A substantial portion of clinically-relevant information for computerized decision support, patient safety, quality improvement and translational research studies is included in narrative text of the Electronic Health Record (EHR) (Jha 2011; Meystre et al. 2008). However, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (Health Insurance Portability and Accountability Act of 1996) requires that before clinical text can be used for research, either (1) all elements of protected health information (PHI) should be removed through the process of de-identification, (2) a patient's consent must be obtained, or (3) the Institutional Review Board should grant a waiver of consent. Studies have shown that requesting consent reduces participation rate, and is often infeasible when dealing with large populations (Dunlop et al. 2007; Wolf and Bennett 2006), for example, thousands of patients. Even if a waiver is granted, documents that include PHI data should be tracked to detect and prevent unauthorized disclosure. On the other hand, de-identification removes the requirements for consent, waiver and tracking, and facilitates clinical NLP research, and consequently, the utilization of information stored in narrative EHR notes.

While manual de-identification on a large scale is all but impossible, NLP can provide a scalable solution. Several studies have investigated the use of NLP to remove PHI elements from clinical narrative texts (Meystre et al. 2010). Rule-based approaches (Friedlin and McDonald 2008; Gupta et al. 2004; Neamatullah et al. 2008) make use of dictionaries and manually designed rules to match patterns of PHI elements. They often lack generalizability and require both time and skill for the creation of the rules, but perform better for rare PHI elements. Machine-learning-based approaches (Deleger et al. 2013, 2014; Gardner and Xiong 2008; Szarvas et al. 2007; Uzuner et al. 2008; Wellner 2009) automatically learn to detect patterns of PHI elements based on a set of examples and are more generalizable, but require a large set of manually-annotated examples. Systems using a combination of both methods usually obtain better results (Meystre et al. 2010; Uzuner et al. 2007).

Overall, the best systems report high recall and precision, often above 90 % and sometimes as high as 99 %. The i2b2 2006 de-identification challenge (Uzuner et al. 2007) provided a benchmark to evaluate and compare a number of automated de-identification systems (Szarvas et al. 2007; Arakami 2006; Guo et al. 2006; Hara 2006). In this evaluation the top systems obtained F-measures around 96 % for overall PHI detection and often superior to 90 % for individual categories, with a few exceptions for categories such as locations (below 80 %) and phone numbers (below 90 %).

Although recent advancements in NLP carry the promise of a highly accurate, automated approach to de-identification, existing de-identification systems present some limitations. These include exclusions of some PHI elements (Uzuner et al. 2008; Fielstein et al. 2004; Ruch et al. 2000), or institution and contact information (Gardner and Xiong 2008; Taira et al. 2002), use of only a limited set of types of clinical note to build and evaluate systems (such as pathology reports (Gupta et al. 2004; Beckwith et al. 2006; Berman 2003)), discharge summaries ((Szarvas et al. 2007; Uzuner et al. 2008; Arakami 2006; Guo et al. 2006; Hara 2006)), or medical message boards (Benton et al. 2011), or use of documents with synthetic PHI elements (manually de-identified documents re-identified with fake PHI elements (Uzuner et al. 2007)). Furthermore, the effect of de-identification and potential over-scrubbing on subsequent automated information extraction tasks is rarely investigated (Meystre et al. 2010).

Recent automated de-identification efforts are moving towards overcoming those limitations. Aberdeen et al. (2010) evaluated their de-identification toolkit (the MITRE Identification Scrubber Toolkit, MIST) on four classes of clinical notes from the Vanderbilt University Medical Center (discharge summaries, laboratory reports, letters, and order summaries) and reported high performance (token-level F-measures ranging 0.934–0.996). They found variations between document classes (highest performance was achieved on the classes which were the richest in PHI elements, such as discharge summaries).

Experiments were also conducted at Cincinnati Children's Hospital Medical Center (CCHMC) to evaluate the MITRE Identification Scrubber Toolkit (MIST) as well as an in-house de-identification system on a larger set of clinical notes representing the variety of notes available in the EHR of a specialized children's hospital (over 20 note types were included in the study) (Deleger et al. 2013). Furthermore, the effect of de-identification on a subsequent IE task (the extraction of medications) was investigated. Evaluation of MIST and the CCHMC system showed performance that was indistinguishable from human annotators: overall precision (P) and recall (R) of the systems were 92.79 % (P) and 92.8 % (R) for MIST and 95.08 % (P) and 91.92 % (R) for the CCHMC system, while precision and recall of the human annotators against the gold standard were 97.68 % (P) and 90.01 % (R) for one annotator and 97.18 % (P) and 98.07 % (R) for the other annotator. Experiments on extracting medications from the de-identified version of clinical notes as opposed to the original version of the notes demonstrated that the automated de-identification caused no significant decrease in the performance of the medication extractor. Results also further emphasized the findings that performance varies depending on note types and their density of PHI elements (Aberdeen et al. 2010).

There are many take-home messages in these recent de-identification experiments that are likely to influence the decisions of Institutional Review Boards regarding whether to accept the output of automated de-identification systems as comparable to manual de-identification. First, no single manual or automated de-identification is 100 % accurate. Second, different note types have different density of PHI elements (and potentially different contexts of the same elements), and a de-identification system that is trained on a mix of note types will show varying performance on these note types. As a result, the de-identification performance of machine learning systems will depend on the frequency of PHI types in the training data. This is not unexpected, but health services researchers and hospital management should be aware of the potential differences in performance for different note and PHI types. In addition, it has been shown that a de-identification system trained on corpora from one institution will only be partially transferable to other institutions (Deleger et al. 2014). Hence, researchers and IT engineers should apply out-of-box de-identification systems more cautiously. Finally, the impact of de-identification seems minimal on the effectiveness of subsequent IE of clinically relevant concepts.

12.3 Phenotype Identification

In May, 2012 the pediatric eMERGE network (Electronic Medical Records and Genomics) was established. Its overarching goal is to combine the phenotype and genotype data into a cohesive EHR, benefiting the patients through appropriate levels of genetic counseling. Success would be a significant advance in field of Electronic Health Record Driven Genomic Research (EDGR) (Kirby et al. 2016; Kohane 2011). One aim of the informatics arm of a pediatric eMERGE project was to demonstrate real-time execution of phenotypic selection across two different pediatric institutions (Children's Hospital Boston and CCHMC) with disparate EHR systems as a model for ensuring phenotypic standardization and for national scalability. Since 2015 the pediatric eMERGE network has entered its third phase, with the focus on integrating exploratory phenotypic algorithms into clinical practice and evaluation.

Pediatric eMERGE's software stack in this case includes the i2b2 platform (Informatics for Integrating Biology and the Bedside (Murphy et al. 2010)) for unifying phenotypic and genotypic information, Apache cTAKES for NLP processing of the clinical narrative, and the distributed query engine called the Shared Health Informatics Research Network (SHRINE) (Weber et al. 2009). The i2b2 software suite integrates the EHR data with clinical research. It has been successfully used to identify adult phenotypes by combining information extracted through NLP with codified data such as lab results and medication orders (Ananthakrishnan et al. 2016; Castro et al. 2015; Liao et al. 2010, 2015; Xia et al. 2013). SHRINE is a general purpose clinical querying protocol that can be adapted to many types of data repositories (i2b2, commercial EHR data warehouses) and has been deployed across multiple institutions (See Chap. 6, Research Patient Data Warehousing). In 2011,

the Catalyst SHRINE went into production at the four main Harvard hospitals for the sharing of aggregate counts of patients with defined exclusion/inclusion criteria for labs, diagnoses, demographics, and medications. CARRAnet, the pediatric rheumatology consortium across 60 institutions, and ICN, a pediatric inflammatory bowel consortium across 40 institutions, worked to test whether i2b2 and SHRINE represent generalizable approaches to federated data sharing among institutions.

The pediatric eMERGE members are linked with the adult eMERGE network that was established in 2007 (eMERGE Network 2012; McCarty et al. 2011). One of the findings is the importance of NLP in achieving the eMERGE goals. As Dan Masys, MD, former Professor and Chair of the Department of Biomedical Informatics at Vanderbilt University, the NIH funded center of the eMERGE initiative, noted in his presentation at an Institute of Medicine workshop: *"Structured data alone (ICD9 codes, labs, meds) is not enough to accurately define clinical phenotypes in EHR data."* and *"Natural language processing of clinician notes improves both sensitivity and specificity of phenotype selection logic* (Institute of Medicine (IOM) 2010)*."*

Numerous publications from the eMERGE research program support the importance of NLP of text of the EHR in phenotyping (for a full list of the publications, consult the *e*MERGE website (2014)). Denny et al. (2010), demonstrated that scanning phenotypes in EHR and grouping patients by EHR-derived phenotypes and analyzing the genetic material of the grouped patients for known mutations is a successful approach to duplicate the findings of Genome Wide Association Studies (eMERGE Network 2012). In the above study Denny et al. used ICD9 codes alone, but in a subsequent publication they proved the importance of NLP for EHR-based phenotyping. Kho et al., reported that "*To assess the additional benefit of NLP, we performed a comparison of the number of cases identified using structured data alone compared with that using both structured data and NLP at one site (Vanderbilt University). At this site, the use of NLP tools identified 129 % more cases of QRS duration (2950 versus 1288) than did the use of structured data and string matching only, while maintaining a PPV of 97 %* (Kho et al. 2011)." Within the adult eMERGE, Kullo and colleagues mine the EHR clinical narrative to discover patient cohorts of confirmed, possible and absent peripheral arterial disease (Savova et al. 2010b). Mo et al. reviewed common features for phenotype algorithms developed within the eMERGE network and presented desiderata for developing a computable phenotype representation model (Mo et al. 2015). They concluded that a computable phenotype representation model should include NLP and text mining, and NLP-driven features have been widely applied for machine learning-based phenotype algorithms.

Another network of researchers focuses exclusively on studying the pharmacogenomics of treatments – the PharmacoGenomics Network (PGRN). Wilke and colleagues (2011) provide an overview of the use of the EHR within PGRN and the role NLP plays in it. EHR data including the clinical narrative processed through NLP has proven accurate for quantifying disease phenotypes and treatment outcomes such as toxicity and efficacy. The NLP-processed clinical narrative combined with medication prescription information was used to develop an algorithm to find the drug treatment patterns a breast cancer patient has undergone (Savova et al.

2011). This cohort was used to study genetic determinants of breast cancer treatments. The Apache cTAKES is used to discover the disease activity of patients with rheumatoid arthritis, build a timeline off it to identify rheumatoid arthritis disease activity (Lin et al. 2013, 2015).

Within the pediatric domain, the team of investigators at CCHMC studied predictive models to determine whether an epilepsy patient is a candidate for neurosurgery (Cohen et al. 2016; Standridge et al. 2014). More than three million people in North America have epilepsy. Nearly 70 % can be controlled with medication. The International League Against Epilepsy guidelines suggest that when seizures cannot be controlled by two or more anti-epileptic drugs, the patient can be considered "medically intractable" and may be a candidate for neurosurgery. The team used all progress notes from 125 epilepsy patients considered medically intractable who had an evaluation for epilepsy surgery, all progress notes from 127 epilepsy patients who were retrospectively determined not to be intractable, and all progress notes from 51 patients whose progress notes were too incomplete to determine their status. Machine learning algorithms were trained and evaluated to achieve an accuracy of 90 percent in distinguishing tractable and intractable epilepsy based on the text in the progress notes. Twenty-four uninformative patients were classified as "maybe non-intractable", 25 as "non-intractable", and two as intractable. A subsequent chart review confirmed that two patients were indeed intractable, as classified by the machine-learning algorithm and should not have been classified as uninformative by medical experts.

12.4 Document and Patient Classification

Depending on the type of the translational and health care research question, the unit of analysis could be one document from the patient's chart or the entire patient's chart. For example, if the overall patient medical history is relevant, then approaches that permit mining the patient's entire medical record are necessary. On the other hand, if the patient's status at a particular time point is of interest, then it is more suitable to mine only specific documents. Thus, document classification is a process to separate a large corpus of diverse documents into smaller homogenous subsets based on predefined criteria (supervised classification) or criteria discovered from the corpus directly without preconceptions of the criteria (unsupervised classification). The most intuitive method for document classification is based on a so-called bag-of-words approach. The documents are converted into vectors that represent unordered collections of words in the documents. In a supervised method, class labels are manually assigned to the training and evaluation gold standard sets (e.g. indicating that one set of radiology reports are consistent with a diagnosis of Acute Lung Injury (ALI), while the second set of documents is not). The vectors are used to train machine-learning algorithms to develop models. Finally, the models are used to classify previously unseen document sets (e.g. radiology reports to assign the ALI label on a large scale automatically) (Solti et al. 2009).

Patient classification could be considered a special case of document classification. Some research questions require the aggregation of information over the patient's entire chart, such as whether the patient is a responder or non-responder to a particular treatment. In the first step of patient classification, all the documents and structured EHR entries are collected for each patient. In subsequent steps, using NLP the information is extracted from the clinical narrative documents and combined with information derived from the structured entries. An aggregate vector representing the patient's chart is assembled. Finally, the vectors are classified with a similar procedure as in document classification. In patient classification the information vectors represent patients instead of documents and the vectors might include information from the structured entries of the EHR, in addition to the documents. Many of the phenotype extractions discussed in Sect. 10.3 are examples of document- and patient- level classifications (Liao et al. 2010; Savova et al. 2010b, 2011).

12.5 Cohort Identification

Many translational and health care studies require the identification of large cohorts of patients who match a set of complex criteria. The implementation of large EHRs has enabled large-scale cohort identification for a variety of investigations – epidemiology, phenotype/genotype associations, biosurveillance. Brownstein et al. (2005), conducted an influenza epidemiology study through monitoring patient visits for respiratory illness using a real-time syndromic surveillance system. Their results added to a growing body of research in support of vaccinating preschool age children. Hansen et al. (2007) mined the EHR data in a large urban academic medical system in northeast Ohio to identify a cohort of 14,187 children and adolescents aged 3–18 to study the underdiagnosis of pediatric hypertension. The following variables were retrieved from the EHR: race/ethnicity, age, sex, weight, height, blood pressure, diagnosis codes, family history of hypertension codes, past medical history codes, problem list codes for each participant from all visits. The investigators concluded that hypertension and prehypertension were frequently underdiagnosed in the studied pediatric population.

Cohort identification from the EHR has been used in studies related to health services research. The American Academy of Neurology (AAN) acknowledges that current evidence is insufficient for robust recommendations for an accurate diagnostic assessment and medical management of children with status epilepticus (SE) (Riviello et al. 2006). SE is defined in the AAN guidelines as a seizure with > 30 min duration that includes two or more sequential seizures without full recovery of consciousness between seizures (Treatment of convulsive status epilepticus 1993). Outstanding translational research questions include (a) whether routine diagnostic testing such as blood cultures, cerebrospinal fluid analysis and neuro-imaging are indicated for the evaluation of these patients and; (b) what the prevalence of acute

bacterial meningitis, significant intra-cranial pathology or bacteremia among these patients is. These are key questions for emergency department physicians, given the risks of radiation exposure (Stein et al. 2008), sedation, and invasive procedures in children. Prospective SE studies have been limited by low incidence (Singh et al. 2010) and low recruitment rates. Analysis of retrospective data has the advantage of the large number of cases potentially available. However, existing data are often limited by the inability to identify a valid cohort of children with SE. Current approaches to data extraction based on chief complaint or ICD9 codes have proven to be rather limited. Such searches normally miss a large proportion of potential cases. Kimia and colleagues (2010) studied a pediatric cohort evaluated in a pediatric emergency department for their first complex febrile seizure. To identify the cohort, the authors created a NLP text screening tool which identified 5744 potential patients out of which the final cohort of 526 eligible patients was selected. Kimia et al.(2009), employed the same approach. They report a sensitivity of 1 and specificity of 0.957 for the NLP text screening tool. The patient cohort was used to study the diagnostic value of lumber punctures for first simple febrile seizure and to assess the rate of bacterial meningitis detected among these children.

The translational work within eMERGE, PGRN and i2b2 projects described in Sect. 10.3 is another example of cohort identification, where the cohort members match phenotypes. Within the adult eMERGE, Kullo and colleagues (2010) mine the EHR clinical narrative to discover patient cohorts of confirmed, possible and absent peripheral arterial disease. Within i2b2, Liao et al. (2010), developed an EHR algorithm to identify cohorts of patients with confirmed, possible and absent rheumatoid arthritis. Patient cohort identification is thus closely tied to phenotype discovery using specific exclusion and inclusion criteria.

One promising area of cohort identification is to automate the matching of clinical trials and eligible patients by learning eligibility criteria directly from the text of trial announcements. This will be the next step in cohort discovery because the textual eligibility criteria descriptions (e.g. on the ClinicalTrials.gov website) will be converted into a computable format, using NLP without or with only minimal human intervention. There are multiple groups working on finding solutions for this task. As an initiative, researchers at CCHMC have developed an automated clinical trial eligibility screener to identify cohorts for pediatric clinical trials (Ni et al. 2015a, b). The automated system uses the cTAKES NLP pipeline and active learning methods to extract: (1) the eligibility criteria compiled from the inclusion and exclusion criteria listed in trial announcements; and (2) the demographics and clinical information from patients' structured EHR data fields as well as unstructured clinical notes. Leveraging information retrieval techniques, it then matched the trial criteria with the patient information to identify potential trial-patient matches. In a gold-standard based retrospective evaluation of 13 pediatric trials, the system achieved statistically significant improvement in screening efficiency and showed the potential of a 91 % reduction in staff screening workload (Ni et al. 2015a). The system was also validated on a set of 55 pediatric oncology trials, where a potential of 85 % reduction in screening effort was observed (Ni et al. 2015b).

12.6 Sentiment Analysis

Who cares? Who doesn't? Why? It is the role of the sentiment expert to find these answers in written text and transcriptions of conversations. Texts filled with sentiment are texts filled with opinions and emotions. They present specific challenges to investigators who wish to extract relevant information from free text, as for example within an EHR. NLP tools are being developed that can support computational approaches to understanding opinions, sentiments, subjectivity, views, and emotions that are found in text. These subjective data are essential to understanding why people make certain decisions (Hu and Liu 2004).

Concepts of opinions and sentiment need to be well defined because opinion mining and sentiment analysis are very different types of tasks. Opinion mining is an identification problem. Opinions are a positive or negative sentiment, view, attitude, emotion or appraisal. Analysis of sentiment is a classification problem. Sentiments are often embedded within opinions and render the opinion positive, negative or possibly neutral. Sentiment analysis is conducted on text, although it can be done on speech as well. Text can be the electronic medical record, blogs, patient portals or short messages. The common thread is that the text must be digitized and not on paper. These texts can be semi-structured or unstructured. Semi-structured texts are ordered and can be assigned to a predefined data model. They have clearly identified headings, sub-headings, tables and figures. As mentioned before, language is usually dubbed as "unstructured information" although any linguist would argue that language does have structure. The term unstructured information is used to distinguish this kind of information from information residing in well-structured databases. Unstructured information does have some level of irregularities and ambiguity which are challenges NLP addresses. Sentiments are usually found in unstructured text (Table 12.1).

In sentiment analysis we want to find the structure in unstructured data (Liu 2011). Then, we have the opportunity to identify patterns, because they provide knowledge (Ackoff 1989). Developing the tools to analyze sentiments is a challenge in translational research and, if successful, could play a valuable role in supporting patient care. In the example of unstructured text in Table 10.1, a clinician first asks: "Who doesn't want to leave her room, Jane or the mother?" If it is Jane, what is the chance that Jane is afraid, abused, shy or just doesn't like the new boyfriend? It is clear that computational analysis will require special algorithms. On the other hand, taking 50 mg of Lamictal is semi-structured. A simple query shows which drug and how much is dispensed. This illustrates the challenge of sentiment analysis, when applied to clinical text.

Table 12.1 Examples of semi-structured and unstructured text

Semi-structured	Unstructured
Current medications: 50 mg of Lamictal	Jane said that since her mother has a new boyfriend she doesn't want to leave her room.

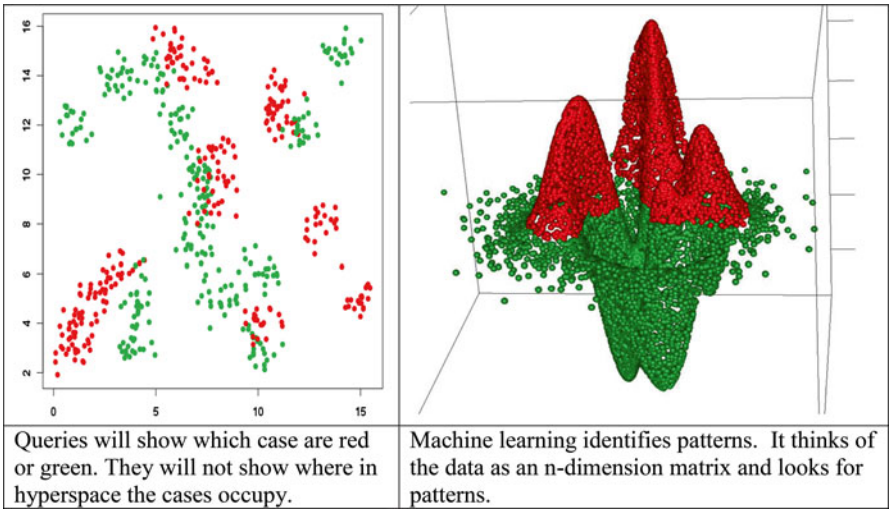


Fig. 12.1 Visualization patterns in 2-D and 3-D space

Sentiment analysis, like most NLP, uses statistical machine learning because these methods can find structures in large data sets (see Fig. 12.1). There are a large number of machine learning algorithms that can be used, but characteristics like sample size, variability and the complexity of the data guide the choice of algorithm (Pestian et al. 2008; Sebastiani 2002).

Sentiment analysis has great potential to support clinical decisions. One example is discovering sentiments embedded in suicide notes, a task requiring NLP technologies. An international competition among over 20 international teams was held to determine which team and algorithm were most accurate in identifying emotions in 1300 suicide notes, compared to a manually annotated gold standard. Their accuracy, as measured by the F_1-measure, ranged from 29 to 61 (Pestian et al. 2012), showing that current tools must be considerably improved to be clinically useful. Pestian et al. (2012), conducted a test of accuracy of sentiment identification by NLP by analysis of suicidal patients. Suicidal and control patients were interviewed with open ended questions, transcripts were prepared, and machine learning was used to analyze the text. Based on the analysis, the origin of the text was assigned to either the control or suicidal group. The accuracy of assignment by machine learning compared to the actual origin (F_1-measure) was 93.3 (Pestian et al. 2012). This shows that it is possible to accurately classify text of notes of suicidal and control patients, using presently available tools of NLP.

The research group also showed that this approach is generalizable. In a novel prospective, multimodal, multicenter, mixed demographic study, machine learning was used to fuse two classes of suicidal thought markers: verbal and non-verbal and classify accurately. Machine learning algorithms that included spreading activation approaches were used with the subjects' words and vocal characteristics to classify 371 subjects recruited from two academic medical centers and a rural community hospital into one of three groups: suicidal, mentally ill but not suicidal, or controls.

By combining linguistic and acoustic characteristics, subjects could be classified into one of the three groups up to 85 % accuracy (Cohen et al. 2015; Pestian et al. 2016; Scherer et al. 2015; Venek et al. 2014).

12.7 "Meaningful Use" of the Electronic Health Record and Comparative Effectiveness

"Meaningful Use" and comparative effectiveness are hot topics widely discussed in healthcare. As part of the Health Information Technology for Economic and Clinical Health Act of 2009, health care providers using a certified EHR are eligible for financial incentives if they meet "meaningful use" objectives (U.S. Department of Health and Human Services (HHS) 2010). Reporting clinical quality measures (CQM) is one such objective. CQMs aim to quantify outcomes and quality of patient care (Centers for Medicare and Medicaid Services (CMS) 2012). Two examples of CQMs relevant to pediatrics are:

- *Measure 0001: Percentage of patients aged 5 through 40 years with a diagnosis of asthma and who have been seen for at least 2 office visits, who were evaluated during at least one office visit within 12 months for the frequency (numeric) of daytime and nocturnal asthma symptoms.*
- *Measure 0002: The percentage of children 2–18 years of age who were diagnosed with Pharyngitis, dispensed an antibiotic and received a group A streptococcus (strep) test for the episode.*

To provide the answers to the above measures, the EHR needs to be queried. Some of the information resides in already structured databases (for example, diagnosis with ICD-9 codes, number of visits, medication prescription), another part is within the clinical narrative (for example nocturnal asthma symptoms, the patient's current medications). In previous subsections, we already introduced one application of NLP, specifically IE. If the clinical narrative has been preprocessed with an NLP-based IE system (for example, cTAKES), a query can then be executed against the SNOMED CT and RxNORM codes generated by the system and used as indices for the clinical narrative. Thus, queries specific to "meaningful use" of the EHR can consume previously processed text.

Comparative effectiveness aims to measure the healthcare outcomes from one approach of disease management to another. As a result, new evidence of the effectiveness of a new intervention or health care service is generated. The Institute of Medicine (IOM) states that "*the overall goal of Comparative Effectiveness Research is the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat and monitor a clinical condition, or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels* (Institute of Medicine (IOM) 2009)." One such study is a report commissioned by the Agency for

Healthcare Research and Quality (AHRQ) on Therapies for Children with Autism Spectrum Disorders conducted by a team of investigators from Vanderbilt Evidence-based Practice Center (Warren et al. 2011). The variables and axes are complex and comprehensive; the authors used publications as their primary source of information which were retrieved through literature searches over PubMED. This study addresses one of the IOM comparative effectiveness priority topics "Compare the effectiveness of therapeutic strategies (e.g., behavioral or pharmacologic interventions, the combination of the two) for different autism spectrum disorders (ASD) at different levels of severity and stages of intervention."

It should be obvious from these examples that one of the main goals of the meaningful use initiative is to facilitate the widespread adoption of informatics in support of ongoing health services and translational research. Another important point to make is that the methods to "query" the EHR to support health services and translation research are not standard methods, yet. The field is under ongoing research and new methods are constantly developed and existing methods refined.

References

Aberdeen J, et al. The MITRE identification scrubber toolkit: design, training, and assessment. Int J Med Inform. 2010;79(12):849–59.

Ackoff RL. From data to wisdom. J Appl Syst Anal. 1989;16(1):3–9.

Ananthakrishnan AN, et al. Identification of nonresponse to treatment using narrative data in an electronic health record inflammatory bowel disease cohort. Inflamm Bowel Dis. 2016;22(1):151–8.

Arakami E. Automatic deidentification by using sentence features and label consistency. In: I2b2 workshop on challenges in natural language processing for clinical sata. 2006.

Aronson, AR, et al. The NLM Indexing Initiative. Proc AMIA Symp. 2000. p. 17–21.

Athenikos SJ, Han H. Biomedical question answering: a survey. Comput Methods Programs Biomed. 2010;99(1):1–24.

Athenikos SJ, Han H, Brooks AD. A framework of a logic-based question-answering system for the medical domain (LOQAS-Med). In: Proceedings of the 2009 ACM symposium on applied computing. ACM: Honolulu; 2009. p. 847–51.

Beckwith BA, et al. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak. 2006;6:12.

Benton A, et al. A system for de-identifying medical message board text. BMC Bioinf. 2011;12(Suppl 3): S2.

Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. Arch Pathol Lab Med. 2003;127(6):680–6.

Brownstein JS, Kleinman KP, Mandl KD. Identifying pediatric age groups for influenza vaccination using a real-time regional surveillance system. Am J Epidemiol. 2005;162(7):686–93.

Cairns BL, et al. The MiPACQ clinical question answering system. AMIA Annu Symp Proc. 2011;2011:171–80.

cancer Text Information Extraction System (caTIES). [cited 2012 March 19]; Available from: https://cabig.nci.nih.gov/community/tools/caties.

cancer.healthnlp.org. Health NLP. [cited 2016 February 18]; Available from: https://healthnlp.hms.harvard.edu/cancer/wiki/index.php/Main_Page.

Castro V, et al. Identification of subjects with polycystic ovary syndrome using electronic health records. Reprod Biol Endocrinol. 2015;13:116.

Centers for Medicare and Medicaid Services (CMS). Clinical Quality Measures (CQMs). [cited 2012 March 19]; Available from: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/ClinicalQualityMeasures.html.

Chapman W, et al. Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp. 2001. p. 105–9.

Choi JD, Palmer M. Getting the most out of transition-based dependency parsing. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. Association for Computational Linguistics: Portland; 2011a. p. 687–92.

Choi JD, Palmer M. Transition-based semantic role labeling using predicate argument clustering. In: Proceedings of the ACL 2011 workshop on relational models of semantics. Association for Computational Linguistics: Portland; 2011b. p. 37–45.

Christensen LM, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. In: Proceedings of the ACL-02 workshop on natural language processing in the biomedical domain – volume 3. Phildadelphia: Association for Computational Linguistics; 2002. p. 29–36.

Cohen KB, Fört K, Pestian J. Annotateurs volontaires investis et éthique de l'annotation de lettres de suicidés. In: Proceedings of the TALN 2015 workshop on ethics and natural language processing. ETeRNAL (Ethique et Traitement Automatique des Langues). Caen; 2015.

Cohen KB, et al. Early identication of epilepsy neurosurgery candidates with machine learning and natural language processing [Submitted for publication]. Biomed Inform Insights. 2016.

Coursera.org [Standford University]. Natural Language Processing. [cited 2012 June 1]; Available from: https://class.coursera.org/nlp/auth/welcome.

Crowley RS, et al. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J Am Med Inform Assoc. 2010;17(3):253–64.

cTakes (Clinical Text Analysis and Knowledge Extraction System). [cited 2012 June 4]; Available from: http://ohnlp.svn.sourceforge.net/viewvc/ohnlp/trunk/cTAKES/.

Deleger L, et al. Building gold standard corpora for medical natural language processing tasks. AMIA Annu Symp Proc. 2012;2012:144–53.

Deleger L, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc. 2013;20(1):84–94.

Deleger L, et al. Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. J Biomed Inform. 2014;50:173–83.

Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Comput Linguist. 2007;33(1):63–103.

Denny JC, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010;26(9):1205–10.

Dunlop AL, et al. The impact of HIPAA authorization on willingness to participate in clinical research. Ann Epidemiol. 2007;17(11):899–905.

*e*Merge network: electronic medical records and genomics. Publications. 2014 [cited 2016 February 18]; Available from: https://emerge.mc.vanderbilt.edu/publications/.

eMERGE Network: electronic Medical Records & Genomics. A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. [cited 2012 March 19]; Available from: http://gwas.net/.

Fielstein FJ, Brown SH, Speroff T. Algorithmic de-identification of VA medical exam text for HIPAA privacy compliance: preliminary findings. In: Fiesch M, Coiera E, Li YCJ, editors. MEDINFO 2004: proceedings of the 11th world congress on medical informatics. IOS Press: Fairfax; 2004. p. 1590.

Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. J Am Med Inform Assoc. 2008;15(5):601–10.

Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp. 2000: p. 270–4.

Friedman C. Towards a comprehensive medical language processing system: methods and issues. Proc AMIA Annu Fall Symp. 1997. p. 595–9.

Gardner J, Xiong L. HIDE: an integrated system for health information DE-identification. In: Proceedings of the 21st ieee international symposium on computer-based medical systems. 2008. p. 254–9.
Guo Y, et al. Identifying personal health information using support vector machines. In: I2b2 workshop on challenges in natural language processing for clinical data. 2006.
Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol. 2004;121(2):176–86.
Hansen ML, Gunn PW, Kaelber DC. Underdiagnosis of hypertension in children and adolescents. JAMA. 2007;298(8):874–9.
Hara K. Applying a SVM based Chunker and a text classifier to the deid challenge. In: I2b2 workshop on challenges in natural language processing for clinical data. 2006.
Haug PJ, et al. Experience with a mixed semantic/syntactic parser. Proc Annu Symp Comput Appl Med Care. 1995. p. 284–8.
Health Information Technologies Research Laboratory (HITRL). [cited 2012 March 19]; Available from: http://hitrl.it.usyd.edu.au/.
Health information Text Extraction (HITEx). HITEx Manual v2.0. [cited 2012 March 19]; Available from: https://www.i2b2.org/software/projects/hitex/hitex_manual.html.
Health Insurance Portability and Accountability Act of 1996 (HIPAA). P.L. 104–191, in 42 U.S.C. 1996.
Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Methods Inf Med. 1998;37(1):1–7.
Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM: Seattle; 2004. p. 168–77.
IBM. IBM – Watson. [cited 2012 April 5]; n.d. Available from: http://www-03.ibm.com/innovation/us/watson/index.html.
Institute of Medicine (IOM). Initial National Priorities for Compartive Effectiveness Research [Consensus Report]. 2009 [cited 2012 March 19]; Available from: http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx.
Institute of Medicine (IOM). The learning healthcare system in 2010 and beyond: understanding, engaging, and communicating the possibilities. [Workshop]. 2010 [cited 2012 June 1]; Available from: http://www.iom.edu/Activities/Quality/VSRT/2010-APR-01.aspx.
Jha AK. The promise of electronic records: around the corner or down the road? JAMA. 2011;306(8):880–1.
JULIE Lab. Jena University Language & Information Engineering Lab. [cited 2012 March 19]; Available from: http://www.julielab.de/.
Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, Prentice Hall series in artificial intelligence. Upper Saddle River: Prentice Hall; 2000. p. xxvi, 934 p.
Kho AN, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med. 2011;3(79):79re1.
Kimia AA, et al. Utility of lumbar puncture for first simple febrile seizure among children 6 to 18 months of age. Pediatrics. 2009;123(1):6–12.
Kimia A, et al. Yield of lumbar puncture among children who present with their first complex febrile seizure. Pediatrics. 2010;126(1):62–9.
Kirby J, et al. An online repository for electronic mdical record phenotype algorithm development and sharing [in press]. J Am Med Inform Assoc. 2016.
Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12(6):417–28.
Kullo IJ, et al. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc. 2010;17(5):568–74.
Lexical Systems Group. Specialist NLP Tools. [cited 2012 June 1]; Available from: http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html.

Liao KP, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken). 2010;62(8):1120–7.
Liao KP, et al. Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. PLoS One. 2015;10(8):e0136651.
Lin C, et al. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. PLoS One. 2013;8(8):e69932.
Lin C, Karlson EW, Dligach D, Ramirez MP, Miller TA, Mo H, et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. J Am Med Inform Assoc. 2015 Apr;22(e1):e151–61. doi:10.1136/amiajnl-2014-002642. Epub 2014 Oct 25.
Lin C, Dligach D, Miller TA, Bethard S, Savova GK. Multilayered temporal modeling for the clinical domain. J Am Med Inform Assoc. 2016 Mar;23(2):387–95. doi:10.1093/jamia/ocv113. Epub 2015 Oct 31.
Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med. 1993;32(4):281–91.
Liu B. Sentiment analysis and opinion mining. In: Paper presented at the twenty-fifth conference on artificial intelligence (AAAI-11 tutorial). San Franciso; 2011. p. 1–99.
Lucene. Apache Lucene Core. [cited 2012 March 13]; Available from: http://lucene.apache.org/core/.
Mack R, et al. Text analytics for life science using the unstructured information management architecture. IBM Syst J. 2004;43(3):490–515.
Manning CD, Schütze H. Foundations of statistical natural language processing. 2nd printing, with corrections. ed. Cambridge, MA.: MIT Press; 2000. xxxvii, 680 p.
McCarty C, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genet. 2011;4(1):13.
Meystre S, Haug PJ. Evaluation of medical problem extraction from electronic clinical documents using MetaMap transfer (MMTx). In: Proceedings of MIE2005 – the XIXth international congress of the European federation for medical informatics. IOS Press; 2005. p. 823–8.
Meystre SM, et al. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. 2010;10:70.
Meystre SM, et al. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008. p. 128–44.
Mo H, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Inform Assoc. 2015;22(6):1220–30.
Murphy SN, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124–30.
National Centre for Text Mining (NaCTeM). [cited 2012 March 19]; Available from: http://www.nactem.ac.uk/index.php.
Neamatullah I, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak. 2008;8:32.
Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc. 2015a Jan;22(1):166–78. doi:10.1136/amiajnl-2014-002887. Epub 2014 Jul 16.
Ni Y, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak. 2015b;15(1):28.
Nielsen RD, et al. An architecture for complex clinical question answering. In: Proceedings of the 1st ACM international health informatics symposium. ACM: Arlington; 2010. p. 395–9.
Online Colleges.net. Standford introducing five free online classes by Anna Schumann. 2012 [cited 2012 June 1]; Available from: http://www.onlinecolleges.net/2012/03/07/stanford-introducing-five-free-online-classes/.
OpenNLP Tools 1.5.0 API: Sentence Boundary Detector. [cited 2012 June 4]; Available from: http://opennlp.sourceforge.net/api/index.html.

Palmer M, Gildea D, Kingsbury P, The Proposition Bank. An annotated corpus of semantic roles. Comput Linguist. 2005;31(1):71–106.
Pestian JP, et al. Sentiment analysis of suicide notes: a shared task. Biomed Inform Insights. 2012;5 Suppl 1:3–16.
Pestian JP, et al. Machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial [in press]. Suicide Life Threat Behav. 2016.
Pestian JP, et al. Using natural language processing to classify suicide notes. AMIA Annu Symp Proc. 2008. p. 1091.
Riviello Jr JJ, et al. Practice parameter: diagnostic assessment of the child with status epilepticus (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology and the Practice Committee of the Child Neurology Society. Neurology. 2006;67(9):1542–50.
Ruch P, et al. Medical document anonymization with a semantic lexicon. Proc AMIA Symp. 2000. p. 729–33.
Savova GK, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010a;17(5):507–13.
Savova GK, et al. Discovering peripheral arterial disease cases from radiology notes using natural language processing. AMIA Annu Symp Proc. 2010b;2010:722–6.
Savova GK, et al. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. J Am Med Inform Assoc. 2012 Jun;19(e1):e83–9. Epub 2011 Dec 1.
Scherer S, et al. Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis. In: Acoustics, speech and signal processing (ICASSP), 2015 IEEE international conference. 2015; p. 4789–93.
Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv (CSUR). 2002;34(1):1–47.
Singh RK, et al. Prospective study of new-onset seizures presenting as status epilepticus in childhood. Neurology. 2010;74(8):636–42.
Sohn S, et al. Classification of medication status change in clinical narratives. AMIA Annu Symp Proc. 2010;2010:762–6.
Solti I, et al. Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches. Proceedings (IEEE Int Conf Bioinformatics Biomed). 2009;2009: 314–9.
Standridge S, et al. The reliability of an epilepsy treatment clinical decision support system. J Med Syst. 2014;38(10):119.
Stein SC, Hurst RW, Sonnad SS. Meta-analysis of cranial CT scans in children. A mathematical model to predict radiation-induced tumors. Pediatr Neurosurg. 2008;44(6):448–57.
Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc. 2007;14(5):574–80.
Taira RK, Bui AA, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. Proc AMIA Symp. 2002. p. 757–61.
Treatment of convulsive status epilepticus. Recommendations of the Epilepsy Foundation of America's Working Group on Status Epilepticus. JAMA. 1993;270(7):854–9.
Tseytlin E, et al. NOBLE – flexible concept recognition for large-scale biomedical natural language processing. BMC Bioinf. 2016;17(1):32.
U.S. Department of Health and Human Services (HHS). Secretary sebelius announces final rules to support 'meaningful use' of electronic health records [News Release]. 2010 [cited 2012 March 19]; Available from: http://www.hhs.gov/news/press/2010pres/07/20100713a.html.
U-Compare. [cited 2012 March 19]; Available from: http://u-compare.org/index.en.html.
UIMA (Unstructured Information Management Applications). Apache UIMA. [cited 2012 June 4]; Available from: http://uima.apache.org/.
Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc. 2007;14(5):550–63.

Uzuner O, et al. A de-identifier for medical discharge summaries. Artif Intell Med. 2008;42(1):13–35.
Venek V, et al. Adolescent suicidal risk assessment in clinician-patient interaction: a study of verbal and acoustic behaviors. In: Spoken Language Technology Workshop (SLT), 2014 IEEE. 2014.
Warren Z, et al. Therapies for children with autism spectrum disorders. Comparative effectiveness review, AHRQ, Number 26. 2011 [cited 2012 June 1]; Available from: http://www.effectivehealthcare.ahrq.gov/ehc/products/106/656/CER26_Autism_Report_04-14-2011.pdf.
Weber GM, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16(5):624–30.
Weiming W, et al. Automatic clinical question answering based on UMLS relations. In: Proceedings of the third international conference on semantics, knowledge and crid. Shan Xi: IEEE Computer Society; 2007. p. 495–8.
Wellner B. Sequence models and ranking methods for discourse parsing. Waltham: Brandeis University; 2009.
Wilke RA, et al. The emerging role of electronic medical records in pharmacogenomics. Clin Pharmacol Ther. 2011;89(3):379–86.
Wolf MS, Bennett CL. Local perspective of the impact of the HIPAA privacy rule on research. Cancer. 2006;106(2):474–9.
Wu S, et al. Negation's not solved: generalizability versus optimizability in clinical natural language processing. PLoS One. 2014;9(11):e112774.
Xia Z, et al. Modeling disease severity in multiple sclerosis using electronic health records. PLoS One. 2013;8(11):e78927.
Yu H, Cao YG. Automatically extracting information needs from Ad Hoc clinical questions. AMIA Annu Symp Proc. 2008. p. 96–100.
Zeng QT, et al. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak. 2006;6:30.

Chapter 13
Network Analysis and Applications in Pediatric Research

Hailong Li, Zhaowei Ren, Sheng Ren, Xinyu Guo, Xiaoting Zhu, and Long Jason Lu

Abstract Networks, where nodes denote entities and links denote associations, provide a unified representation for a variety of complex systems, from social relationships to molecular interactions. In an era of big data, network analysis has been proved useful in biological applications such as predicting functions of proteins, guiding the design of wet-lab experiments, and discovering biomarkers of diseases. Driven by the availability of large-scale data sets and rapid development of bioinformatics' tools, the research community has applied network analysis to define underlying causes of pediatric diseases. This will almost certainly lead to more effective strategies for prevention and treatment of diseases. In this chapter, we will

H. Li
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, USA

Z. Ren • X. Guo • X. Zhu
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, USA

Department of Electrical Engineering and Computing Systems, University of Cincinnati, 2600 Clifton Avenue, Cincinnati, OH 45221, USA

S. Ren
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, USA

Department of Statistics, University of Cincinnati, 2600 Clifton Avenue, Cincinnati, OH 45221, USA

L.J. Lu (✉)
Department of Electrical Engineering and Computing Systems, University of Cincinnati, 2600 Clifton Avenue, Cincinnati, OH 45221, USA

Department of Environmental Health, University of Cincinnati, 2600 Clifton Avenue, Cincinnati, OH 45221, USA

Departments of Pediatrics and Biomedical Informatics, Division of Biomedical Informatics, Cincinnati Children's Hospital Research Foundation, 3333 Burnet Avenue, MLC 7024, Cincinnati, OH 45229, USA

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, USA
e-mail: long.lu@cchmc.org; http://dragon.cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_13

introduce classic and the state-of-the-art network analysis methodologies, approaches and their applications. We then provide four examples of recent research, where network analysis is being applied in pediatrics. These include the identification of high-density lipoprotein particles that underlie the development of cardiovascular disease using protein-protein interaction networks, alternative splicing analysis by splicing interaction network, construction and network analysis of pediatric brain functional atlas, and disease relationship exploration using diagnosis association networks constructed by electronic health record.

Keywords Network analysis • Network applications • Network prediction • Pediatric diseases

13.1 Introduction

Networks, or graphs, provide a unified representation for a variety of complex systems, from social relationships, e.g., co-authorship of different authors, to associations between molecular entities. Molecular networks, where each node denotes a molecule like a protein or gene and each edge denotes an association, form the foundation of contemporary systems biology. To date, network models are applied to depict a variety of biological systems. For example, protein-protein interaction (PPI) networks or maps, where each node is a protein and each edge is an interaction, are adopted to describe physical interactions or attachments between protein pairs in sets of proteins such as those found within a cell organelle, or particle (Rual et al. 2005). Gene co-expression networks, where each node is a gene and each edge indicates gene expression similarity, are applied to summarize similarities in expression patterns between gene pairs (Stuart et al. 2003). Transcriptional regulatory networks, where each node is either a transcription factor protein or a target gene, depict transcriptional regulatory relationships between transcription factors and target genes (Guelzim et al. 2002). Other common biological networks include metabolic networks of metabolites and their chemical reactions and signal transduction pathways of multiple interacting genes and proteins (Zhang et al. 2010). Analyzing these large-scale networks may provide a system-level view of these biological systems and thus advance understanding of the underlying biological processes. In recent years, applications of network analysis have been proven useful in guiding experimental design, uncovering novel and effective prognostic biomarkers, and facilitating drug discovery. Some basic features of networks are illustrated in Fig. 13.1.

In this chapter, we will first introduce tools and procedures for state-of-the-art biological network analyses and provide a few examples of their application to basic research on causes of pediatric diseases. Next, we will introduce the global properties of large-scale biological networks: scale-free topology, small-worldness, disassortativity, and modular structures. In the third section, we will discuss in brief

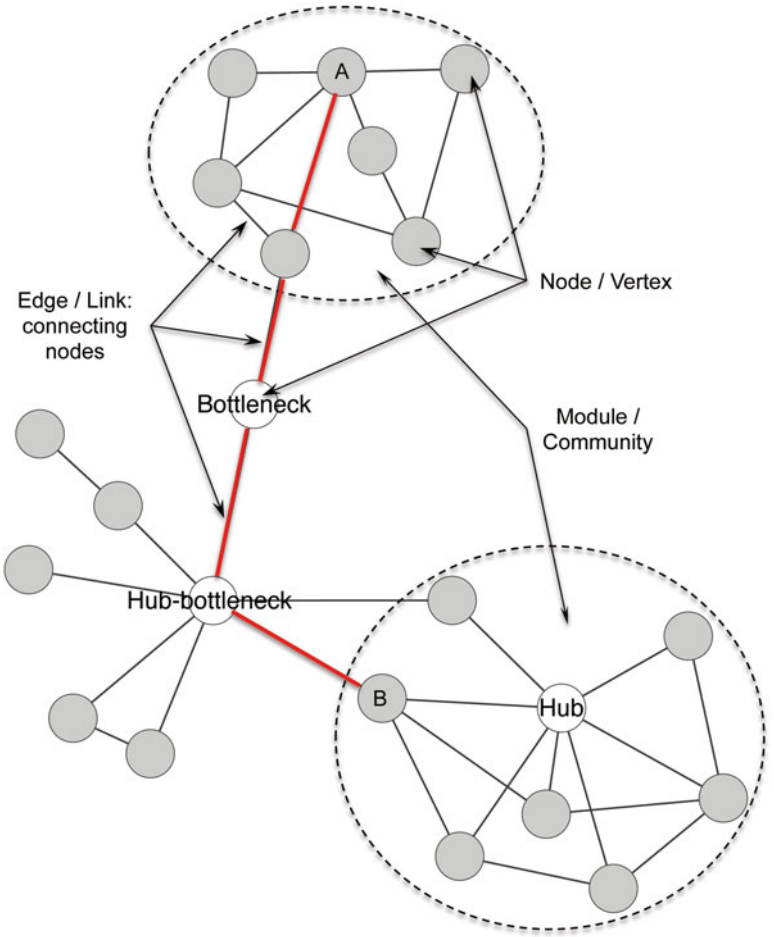


Fig. 13.1 Basic features of networks. Nodes or vertices are entities in a network, and edges or links connect pairs of nodes. The *red* edges form a path that connects nodes A and B. A hub node is defined as a node with a large number of neighbors, and a bottleneck is defined as a node that many shortest paths go through. In this sample network, the node in the *lower left* is both a hub and a bottleneck. Fifteen nodes in a network with 22 nodes form two different densely intra-connected and loosely inter-connected modular subnetwork structures called modules or communities (in *dashed ellipses*)

four common types of network analyses and show how they have been applied: (1) topological analysis to identify proteins/genes of interest, (2) motif analysis to extract and identify small subnetwork motif structures, and (3) modular analysis to cluster large molecular networks to self-contained subnetworks as functional units or modules. Finally, examples of analyses will be provided, including (1) analysis of proteins and protein-protein interactions in subspecies of high density lipoprotein particles that play a role in cardiovascular disease, (2) alternative splicing analysis

using a splicing interaction network by AltAnalyze, (3) construction of brain functional connectivity maps from neuroimaging data in children with a focus on attention-deficit/hyperactivity disorder (ADHD), and (4) construction of diagnosis association network by electronic health record. Chapter 16 (Bioinformatics and Orphan Diseases) and Chap. 19 (Functional Genomics – Transcriptional Networks Controlling Lung Maturation and Surfactant Homeostasis) provide additional examples of the application of network analysis to studies of childhood diseases.

13.2 Biological Networks: Properties and Characteristics

Large-scale biological networks of various types are similar in a number of topological properties. They are scale-free, small-worldness, disassortativity, and modular structures. We will illustrate each of them one by one.

Similar to many other large-scale networks such as the World Wide Web and social networks, most large-scale biological networks of model organisms are observed to approximate a scale-free structure. Scale-free means that, for a node in the network, the probability distribution of the number of connections to other nodes (degree distribution) follows a power-law distribution, denoted as $P(k) \sim k^{-\gamma}$, where $P(k)$ is the fraction of nodes that have k connections to other nodes (degree k), and the degree exponential constant γ is a constant usually smaller than 3 (Barabasi and Oltvai 2004). Figure 13.2 shows the scale-free topology of a large-scale human protein-protein interaction (PPI) network, with the PPI data from the Human Protein Reference Database (Goel et al. 2012). Scale-free topology implies that only a small fraction of nodes have a very high degree of connections (large number of connections), and those nodes are called hubs. Hubs are critically important to the overall functioning of the network. If there is an attack targeted at hubs, the network could be easily destroyed. However, compared to random network, the scale-free network are robust to random attack. For example, one study showed that removing up to 80 % of randomly selected nodes in a scale-free network would not disconnect the rest of it, due to the sparseness of connections for most of the nodes (Albert et al. 2000).

Most large-scale biological networks have small-world property as a consequence of their scale-free topology. Small-worldness can be defined as a network with high clustering coefficient and low characteristic path length, which means that nodes in network are involved in community-like subnetworks and most node pairs are connected via short paths (Fig. 13.1). For example, in a *Escherichia coli* metabolic network of 287 nodes and 317 edges, where a node is a metabolite and an edge represents a reaction, a typical path between the most distant metabolites contains only four reactions (Wagner and Fell 2001). The small-world property also contributes to the robustness of a network, because when any perturbation occurs, most of the network components can conduct a timely response because of the small-worldness.

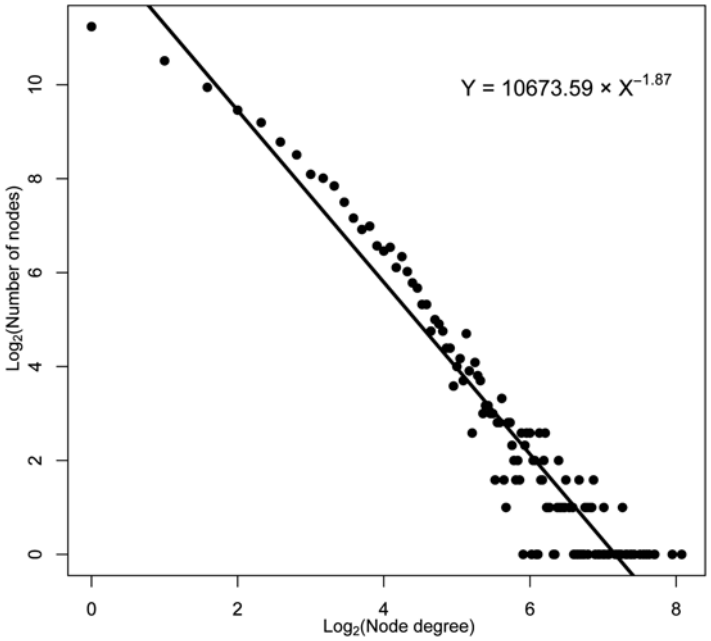


Fig. 13.2 The scale-free topology of a large-scale PPI network. The node degree of a large-scale PPI network based on the 9th release of HPRD follows a power-law distribution (a *straight line* after log transformation on both nodes number and nodes degree). A total of 9517 proteins and 37,004 pairwise PPIs exist in the network

Disassortativity, which means hub and non-hub nodes are more likely to be connected than randomly expected, was discovered in large-scale PPI as well as transcriptional regulatory networks (Maslov and Sneppen 2002). The disassortativity, in accord with scale-free and small-world characteristics, strengthens the robustness of the network under random attack. This property also helps to separate subnetwork structures in a network, which may correspond to functional units that perform particular functions in biological processes.

Large-scale networks are composed of functional modules that perform distinct, but relevant, functions (Hartwell et al. 1999). In biological networks, such functional modules correspond to modular subnetwork structures, which are highly intra-connected within themselves and loosely inter-connected with each other. The molecules within each module are likely to be annotated with the same or similar function. In most large-scale biological networks, the modular structures are often hierarchically organized, due to hierarchical modularity of biological networks and scale-free property (Ravasz 2009).

13.3 Network Analysis and Applications: Topology, Motifs and Modules

The most common types of network analysis include topological, motif, and modular analysis. In this section, we introduce these methods of network analysis and their applications to knowledge discovery. Table 13.1 lists some online tools/software commonly used for network analysis together with their function and how to access them. Typical outputs of these analyses are shown in Figs. 13.1, 13.5, 13.7 and 13.8.

Network topology is often the first thing to consider given a biological network. In order to quantitatively measure the topological structure of a network, a number of classic network statistics from graph theory have been adopted. Commonly used network statistics include degree, centrality, clustering coefficient, average/shortest path length, and graph eccentricity. These network statistics not only describe single nodes properties in the network, but also characterize the network as a whole.

In an undirected network such as a PPI network, node degree is defined as the number of connections linked to a node. In a directed network such as a transcriptional regulatory network, the number of out-going and in-coming links of a node can be measured separately as out-degree and in-degree. The average degree of all nodes in a network is an indication of how dense or sparse the connections are. Similar to other real-life networks, biological networks are sparse, which is defined as having far fewer than the possible maximum links $\frac{n\times(n-1)}{2}$, where n is the number of nodes. The clustering coefficient of a node can be defined as the fraction of observed links compared to all possible links between its neighboring nodes. A node with clustering coefficient close to 1 implies that its neighbors are close to becoming a clique (complete graph). The average clustering coefficient of all nodes in a network is an indicator of whether they are involved in densely connected subnetworks. The higher the average clustering coefficient of a network is, the more likely that its nodes form self-contained subnetwork modules. The shortest path length between two nodes is the minimum number of links needed to connect two nodes. Characteristic path length for a network is defined as the median of the means of the shortest path lengths of all nodes taken together. Small-worldness is defined as having high clustering coefficient and low characteristic path length.

A network can be analyzed at node level. In biological networks, two important kinds of nodes – hubs and bottlenecks, are of special interest. A hub is an influential local connector, while a bottleneck is a bridge-like connector between different communities (Fig. 13.1). A number of studies have reported enriched functional essentiality in hub or bottleneck nodes of protein-protein interaction networks of model organisms, or high correlation between topological connectivity and functional essentiality (Jeong et al. 2001; Yu et al. 2007). Based on these observations, one application of retrieving hubs and bottlenecks is to identify important proteins or genes in biological networks. Hubs or bottlenecks in the PPI networks of human pathogens may help identify essential proteins for the survival of these pathogens.

Table 13.1 Examples of tools (software applications) commonly used in network analysis

Tool	Function	How to access
Cytoscape. Shannon et al. (2003)	Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data. Cytoscape supports many use cases in molecular and systems biology, genomics, and proteomics. Various plug-ins for analysis and visualization have been developed in addition to its core functionalities	www.cytoscape.org/ (Last update in 2016)
Gephi. Bastian et al. (2009)	Gephi is an open source standalone software platform for visualizing networks and complex systems. Gephi provides topological and modular analysis for biological and social networks. A variety of plug-ins are also available for analysis, visualization layout, and applications of specific purposes	www.gephi.org (Last update in 2016)
VisANT. Hu et al. (2009)	VisANT provides a webstart application, a standalone Java application, and a batch mode for visualization and analysis. It has multiple layout options and is scalable to visualize genome-scale biological networks. Users can perform statistical, motif, and modular enrichment analysis using VisANT	visant.bu.edu (Last update in 2015)
Pajek. Batagelj and Mrvar (2004)	Pajek is a standalone application tool for network visualization. Pajek supports 3D layout and can perform modular analysis of networks to decompose them into modules	http://mrvar.fdv.uni-lj.si/pajek/ (Last update in 2016)
tYNA. Yip et al. (2006)	tYNA is a Web server that for biological network visualization. tYNA supports basic network analysis of major topological characteristics and motifs	tyna.gersteinlab.org/
JUNG. O'Madadhain et al. (2003)	JUNG, short for Java universal network/graph framework, is a Java-based open-source software library package. JUNG provides various implemented algorithms for network modeling, visualization and analysis. It is a general library and is not specific to biological network applications	jung.sourceforge.net/ Newest release version (2016): https://sourceforge.net/projects/jung/

(continued)

Table 13.1 (continued)

Tool	Function	How to access
PINA. Wu et al. (2009)	PINA is a Web server primarily for protein-protein interaction network visualization and statistical analysis. Important nodes such as hubs and bottlenecks can be identified. PINA can perform GO term enrichment analysis. It has good scalability	cbg.garvan.unsw.edu.au/pina/ (Last update in 2014)
N-Browse. Kao and Gunsalus (2008)	N-Browse is both a webstart application and a Web-server for visualizing molecular networks. It is connected to and integrated with multiple databases such as modMine database for biological data mining	http://aquila.bio.nyu.edu/NBrowse2/NBrowse.html

Such proteins could be potential targets for developing drugs to treat infections caused by those pathogens.

A network can be further analyzed at motif level. A network motif is defined as an interconnected subnetwork structure of several nodes that occurs much more frequently than random expectation. Motifs have been identified in a wide range of networks, such as transcriptional regulatory networks, social networks, electrical circuits, the World Wide Web, and ecological food webs. Figure 13.3 shows several types of commonly identified motifs in directed networks. Despite different sizes, types, and complexity of networks, the approaches to identify network motifs are generally similar. A common method involves scanning a network for all patterns with several interconnected nodes, recording frequencies of their occurrence, and comparing pattern frequencies with those in networks that have randomly rewired links. A subnetwork pattern is considered a motif, if it has a much higher frequency than randomly expected, often defined as two or more standard deviations greater than the average number of occurrences in random networks. Different types of motifs are over-represented in networks with different types of functions. For example, in transcriptional regulatory networks, the "feed-forward loop" motif is common.

Moreover, modular analysis of a network could be more important, especially for biological networks which are highly modular. Structurally, they are composed of densely interrelated nodes of proteins or genes. These subnetworks have high connectivity density and are loosely interconnected (Fig. 13.1). Functionally, modules in biological networks are those interconnected molecules that together perform specific functions. Structural modules are believed to correspond to functional modules. In the past decade, the research community has made efforts to uncover functional modules through the extraction of structural modules from large-scale biological networks. Various graph clustering algorithms and methods have been developed or adopted for modular analysis. These methods can be roughly categorized into four types based on the underlying methodologies, they are density-based,

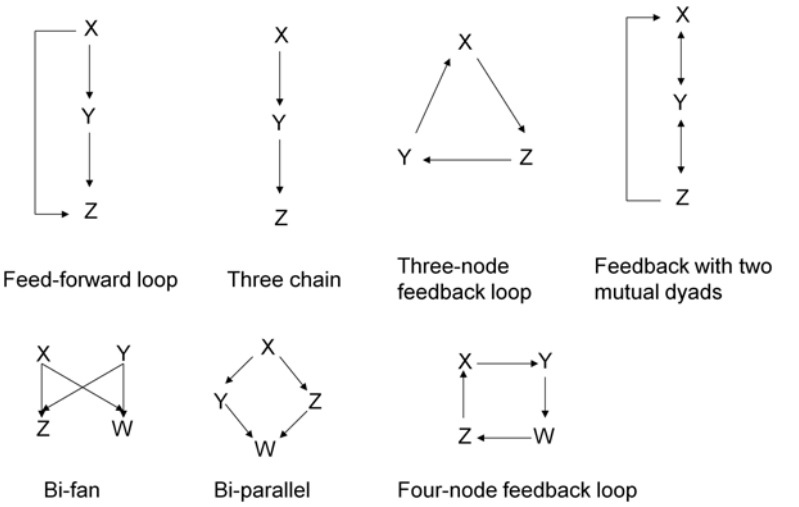


Fig. 13.3 Commonly identified motifs in directed networks. Each letter "x", "y", "z", or "w" indicates a node. Each motif is consisted of 3–4 nodes in these commonly identified motifs

partition-based, centrality-based, and hierarchical clustering approaches (Zhang et al. 2010). Many of them are implemented in the tools listed in Table 13.1.

A density-based clustering method seeks to identify densely connected subnetwork structures in a network, for example fully connected or near-fully connected subgraphs. Clique enumeration is a simple and straightforward density-based method, which can identify all cliques of up to *n* nodes (Spirin and Mirny 2003). Other density-based methods may seek to find densely connected subnetworks with less stringent connectivity criteria than fully connected cliques. For example, *Palla et al* developed a clustering method to identify *k*-clique subnetwork structures, which are defined as unions of all cliques of *k* nodes that share *k*–1 nodes (Palla et al. 2005). In a biological network such as a PPI network, proteins in a clique or clique-like module likely correspond to those belong to the same protein complex. Modules identified by density-based methods can overlap with each other. Such a property is favored in biological networks because proteins and genes may participate in different functional units under different contexts. One limitation of density-based methods is that a module may never cover loosely connected nodes in a network.

Unlike density-based clustering methods, partition-based methods seek to identify an optimal partition of all nodes in a network based on a certain cost function, which measures how a particular partition fits the data. Such a method often starts with a random partition of a network, followed by iterative reassignment of nodes from different previous partitions until a minimal cost is achieved. In each iteration, the cost function is evaluated, and the method outputs a partition when a minimal cost is achieved. One example of such cost function is "modularity", a score between 0 and 1, where a higher score indicates many more within-module links and many

less between-module links than expected. A limitation of such methods is that network partitioning assigns nodes to distinct modules and does not allow overlapping modules.

Centrality-based clustering methods rely on network centrality statistics. One commonly used statistic is edge/node betweenness, which is defined as the number of shortest paths (edges that are connected by nodes) between all possible node pairs that pass through an edge/node. Edges/nodes with high betweenness correspond to bridge-like connectors in a network, and the disconnection of which would result in self-contained node clusters. A classic edge betweenness based clustering method takes all the links in a network as input, and iteratively removes an edge with the highest edge betweenness (Girvan and Newman 2002). A network can be clustered into subnetwork modules after a certain percentage of edge removal. In a biological network, the resulting modules from a centrality-based clustering method often form a hierarchy because of the hierarchical modularity properties of biological networks.

Classic hierarchical clustering methods have also been applied to biological networks clustering. In a PPI network, distance measures between pairs of proteins in a network can be defined by their pairwise distance (shortest path length d) or a function of pairwise distance, e.g. d^2. The proteins are grouped together from the closest in the network to the farthest. The grouping forms a hierarchy, and given a cutoff of the distance, proteins can be put into different clusters. One limitation of hierarchical clustering is that different linkage methods and different cutoffs would result in different output modules. In addition, classic hierarchical clustering algorithms do not consider overlapping clusters.

Predicting the function of proteins is also a major application of network modular analysis in PPI networks. In PPI networks, subnetworks identified via network clustering are likely corresponding to functional modules. Despite the variety of clustering methods and algorithms, protein members in each subnetwork tend to have the same or similar functional annotations, e.g., Gene Ontology annotations (Berardini et al. 2010). Based on PPI network clustering results, proteins with unknown functions can be predicted to have functional annotations that are highly enriched for the groups they belong to (Brun et al. 2004; King et al. 2004).

13.4 Applications of Network Analysis

13.4.1 Protein-Protein Interaction Networks in High Density Lipoprotein Particles

Network analysis can be applied to characterize protein-protein interactions (PPIs) in complex mixtures of proteins within subcellular particles. In our recent study on high density lipoprotein (HDL), a network-based approach was employed to systematically infer the HDL subspecies. We identified and characterized structural HDL subspecies through analysis of proteins' co-migration patterns generated by

three orthogonal chromatograph separation techniques using a multi-network-based approach rather than previous single HDL PPIs (Li et al. 2015).

HDLs are blood-borne complexes consisting of proteins and lipids that play critical roles in cardiovascular disease (CVD) (Boden 2000; Lewis and Rader 2005; Cuchel and Rader 2006). There is growing controversy about how HDL prevents CVD. A widely accepted mechanism is called reverse cholesterol transport, where HDL promotes cholesterol efflux from peripheral cells such as macrophage-derived foam cells in the vessel wall to transport excess cholesterol and other lipids back to the liver for catabolism (Franceschini et al. 1991). HDL is also involved in other CVD-protective functions, including anti-oxidation, anti-inflammation and endothelial relaxation (Watson et al. 1995; Naqvi et al. 1999; Nofer et al. 2002; Barter et al. 2004; Negre-Salvayre et al. 2006; Mineo et al. 2006).

There is evidence that HDL is composed of numerous distinct particle subpopulations, each containing a unique protein makeup that plays critical physiological roles (Gordon et al. 2010a, b). The major activities of HDLs rely on the cooperative interactions among its protein components. Recent proteomic studies on HDL have identified upwards of 89 distinct HDL associated protein components (Gordon et al. 2010b; Heller et al. 2005; Karlsson et al. 2005; Rezaee et al. 2006; Vaisar et al. 2007; Davidson et al. 2009). We developed three non-density based orthogonal separation chromatography techniques, including gel filtration chromatography (GF), anion exchange (AE) chromatography, and isoelectric focusing (IEF) chromatography, to fractionate HDL particles by molecular size, charge and isoelectric points respectively, and high-performance liquid chromatography/electrospray ionization tandem MS (HPLC-ESI-MS/MS) was used to identified the proteins within subpopulations (Gordon et al. 2010b). GE separated each plasma sample into 17 successive size-based fractions across 106 identified lipid associated protein. AE identified 140 proteins in 21 fractions. IED identified 93 proteins in 20 fractions. We compared the existing data from 16 proteomics studies published to date and compiled a list of 89 high-confidence HDL-associated proteins that were observed in at least three different studies and have independent.

To identify HDL subspecies, individual co-migration networks were constructed for each fraction from the different separation techniques (Fig. 13.4). Each vertex represented a HDL-associated protein and edges represented the local co-migration relationship. We developed the local Spearman's rank correlation coefficient score (local S-score) to quantitatively measure the co-migration relationship between any pair of HDL proteins as local similarity. 58 protein local co-migration networks were constructed and merged into a comprehensive HDL interactome map containing 70 proteins and 1540 edges to illustrate the overall interactions within HDL proteome.

A maximal clique identification method called *Bron-Kerbosch* algorithm was applied within each co-migration network to discover HDL subspecies from core network elements. As all proteins in one subspecies have similar migration patterns, they tend to form a highly connected cohesive network module that can be represented by a maximum clique, which is a fully connected subset of the vertices that every two nodes within are connected by an edge and including any more adjacent

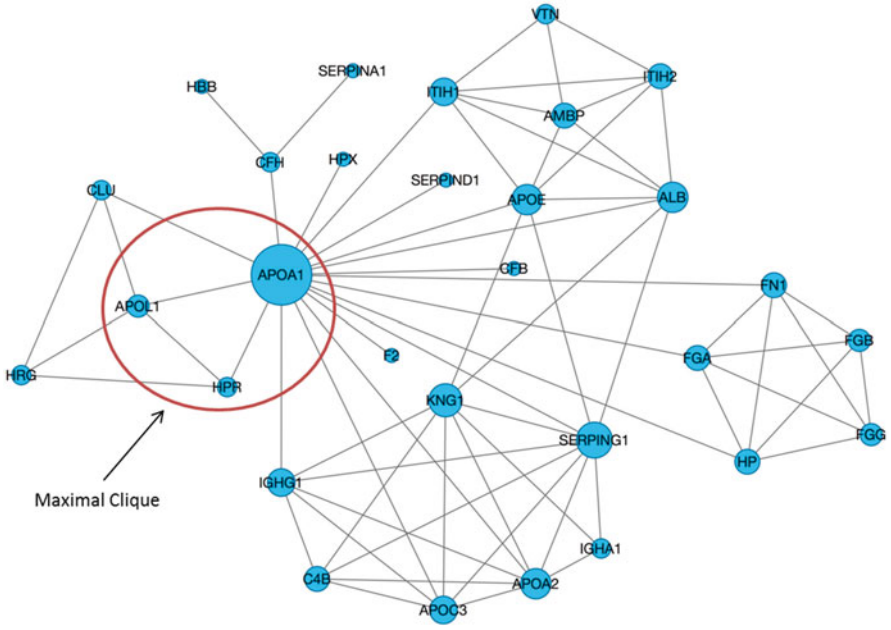


Fig. 13.4 A local co-migration network constructed for 19th fraction of GF method. Size of the vertex reflects network degree of this vertex. *Circled* subnet is a maximal clique within this network, corresponding to the well-known TLF particle

nodes. 183 candidate cliques containing at least three nodes were selected after filtering by molecular weight. Six lines of evidences were used to test the validation of the identified subspecies, including literature analysis, GO function analysis, two PPIs models based on our mouse HDL study and two PPIs models based on our human genetic deficiency disease study, and 38 of the candidate cliques are further selected as potential HDL sub-particles.

We compared our local-network approach with traditional methods on network construction and clique identification. We assessed the accuracy of the predicted HDL interactome network based on known HDL PPIs deposited in the Human Protein Reference Database (HRPD) with sliding window, and illustrated that out local-network model has better prediction power that traditional model. We also constructed three PPI networks from three separation methods and only considered the overlapping edges and found only 5 of 352 well-known PPIs with negligible coverage, suggesting that our local-network method was tolerant of potential artifacts due to experimental perturbation.

We elucidated the functions of each clique by performing functional enrichment analysis on these cliques based on GO annotations. Of the 38 subspecies candidates, 31 have significant enriched functions after Benjamin correction (p-value < 0.01). The most enriched functions covered previously known HDL functions include "reverse cholesterol transport", "anti-oxidation", "immune response", and "hemostasis", as well as essential functions to keep the normal cellular activities such as

"positive regulation of heterotypic cell-cell adhesion". These results are consistent with known knowledge of HDL functions but suggest that subpopulations of the HDL particles differ in functional profiles. Furthermore, gene knockout experiment in mouse model supported the validity of these sub-particles related to three apolipoproteins (Gordon et al. 2015). Analysis of an apoA-I deficient human patient's plasma provided additional support for apoA-I related complexes.

13.4.2 Splicing Interaction Network Application by AltAnalyze

Abnormal RNA splicing has been attributed to as much as 60 % of common and sporadic genetic diseases (Xiong et al. 2015; Zhang et al. 2015; Hull et al. 2007; Wang and Cooper 2007). Many pediatric diseases specifically involve mutations that directly or indirectly impact alternative splicing, through mutation of a splicing factor or splicing factor binding site. These include congenital heart diseases (Kerstjens-Frederikse et al. 2016; Homsy et al. 2015; Lu et al. 2016), immunological defects (Koh et al. 2015; Chen et al. 2015; Lucas et al. 2014) and neurodevelopment (Schaffer et al. 2014). In the case of mutations that impact splicing factors, such as CLP1 in neurodevelopment or RBFOX in congenital heart disease, a network of mis-spliced RNA products is likely to underlie the disease phenotypes. A number of network-based approaches are now under active development to infer these splicing networks and understand their involvement in larger physiological pathways. Recent experimental data has highlighted the importance of isoform specific interactions and it is likely to contribute to pediatric diseases (Corominas et al. 2014; Yang et al. 2016). Here we will introduce the software AltAnalyze and its Cytoscape plugin DomainGraph to analyze RNA-Seq or raw microarray data and visualize genes in splicing networks (Emig et al. 2010). It provides a comprehensive solution to perform alternative splicing analysis on microarray data and visualization of splicing and interaction network. It focuses at the level of proteins, domains, microRNA binding sites, molecular interactions and pathways.

AtlAnalyze performs alternative exon and functional prediction analyses. When a user is interested in interactions among genes in a particular pathway or gene set, AltAnalyze can be used to comprehensively evaluate protein isoforms and protein interactions. DomainGraph is a downstream component of AltAnalyze for visualizing biological effects of alternative spliced genes. The statistics of gene and pathway information from AltAnalyze are used as input for DomainGraph. Users can import either gene or protein interactions into DomainGraph. When gene interactions are imported, the focus lies on isoforms and whether they are affected by alternative splicing. In contrast, when protein interactions are imported, the focus lies on interactions between protein isoforms. If a dataset shows significant up- or down- regulated genes, biological information (e.g. gene symbols, pathways) are automatically generated. The biological dataset could be visualized along with the isoform information (gene, transcript, exon, protein, and Pfam domains) and expression level. Users are allowed to start the analysis in AltAnalyze, and load any

interesting gene, pathway or network into DomainGraph. Then the effects of alternative splicing are comprehensively explored and evaluated at any level (from network-level to exon-level).

When AltAnalyze is used to analyze an Affymetrix microarray, the input data is a CEL format microarray data. The result of AltAnalyze is a gene list with p-value of alternative exons with each gene. Then data are imported to DomainGraph and analyzed by calculating enrichment on pathway, annotation and alternative splicing event. We selected an enriched pathway 'Apoptosis' and load it in DomainGraph in pathway view (Fig. 13.5). Interaction and pathway information were obtained from public pathway database (e.g. WikiPathways (Kutmon et al. 2016) and Reactome (Croft et al. 2010)). WikiPathways and Reactome are both free, curated and peer reviewed pathway database. In the pathway view, each gene is represented by a rectangle node, and genes are connected with directional edges to display the interactions between them. Yellow boxes represent those genes in the input list with alternative exons.

AltAnalyze is also easy to cooperate with other sequencing data or visualization software. Meta-Analysis of Affymetrix Microarray Data analysis (MAAMD) (Gan et al. 2014) is a software to perform meta-analysis on microarray data. AltAnalyze is embedded into MAAMD as a tool for splicing analysis. In another computational workflow for whole genome RNA-Seq analysis (Soreq et al. 2014), an updated version of AltAnalyze is combined to directly evaluate differential gene expression,

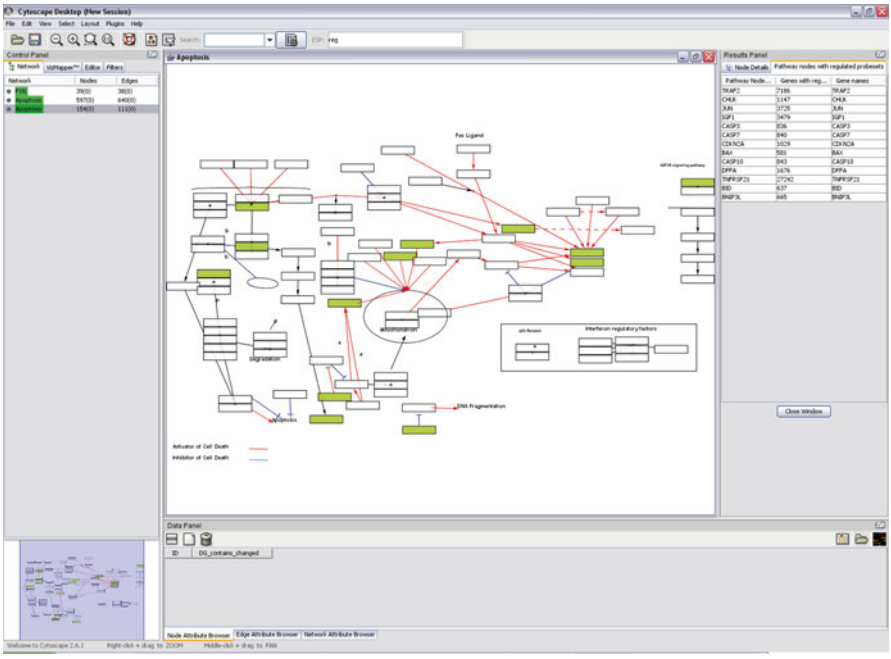


Fig. 13.5 Pathway view of 'apoptosis' pathway in DomainGraph

identify known and novel alternative exons and junctions and alternative Poly-A sites, perform combinatorial exon and junction analyses and evaluate these effects at the level of protein domains, miRNA targeting and enriched biological pathways, as a fully automated and user-friendly pipeline.

In sum, AltAnalyze and DomainGraph are capable of providing a visualized way to study biological impact of alternative splicing on a pathway or network view. They provide several novel functions, including finding miRNA binding sites, protein domain analysis and interaction networks integrated with domain and isoform information from online public database. The analysis focuses on mammalian species at genome scale. The statistical analysis on alternative splicing event from AltAnalyze identifies the significance of potential alternative splicing events. The visualized pathway network analysis from DomainGraph provides an interface to comprehensively understand the biological effects on specific group of genes and isoforms.

13.4.3 Detection of Brain Functional Connectivity Map in Children with Attention-Deficit/Hyperactivity Disorder (ADHD)

Imaging techniques, such as magnetoencephalography (MEG), positron emission tomography (PET), and magnetic resonance imaging (MRI), provide opportunities for analyzing both structural and functional organization of human brains. Traditional analysis of brain images focuses on the study of individual brain regions. The differences between the brains of two groups of humans (e.g., normal versus disease) are compared on a region-by-region basis. There are limitations to this approach. First, some disorders may affect a large number of brain regions, and the significance of changes in any individual region is difficult to establish. Second, the functional impairment may be caused by abnormalities in the interconnections among several brain regions. To overcome these limitations, the different anatomic regions of the brain must be analyzed as an interconnected whole.

Network analysis takes into account patterns of connections among brain regions or voxels in digital images. Different from molecular networks presented in the previous sections, the nodes in brain networks represent voxels or brain regions, depending on the expected granularity of the analysis, and the edges represent the structural or functional connections among the different voxels or brain regions.

There are three main types of brain networks: structural networks, functional networks, and effective networks. The edges in these three brain networks represent physical connections, functional correlations, and causal functional relations, respectively. Brain networks can be represented as weight matrices or graphs. The forms of representation for the structural, functional, and effective networks are similar to one another. Topological analysis can be performed on the brain networks to reveal the connection patterns of different groups of brains, e.g., healthy versus diseased. The results of analysis can reveal the topological differences that cannot

be revealed by traditional non-network based approaches. The topological differences can provide new insights into the mechanism of the diseases.

Attention Deficit Hyperactivity Disorder (ADHD) is a psychiatric disorder characterized by clinical symptoms of inattention, impulsivity, and hyperactivity. This condition affects 5–8 % of school age children, and usually persists into adolescence and adulthood. Clinical diagnosis of ADHD is based on behavioral information gathered from parents and school. Depending on the number and type of symptoms, a child can be diagnosed with one of three ADHD presentations: primarily inattentive (ADHD-I), primarily hyperactive (ADHD-H) or combined subtype (ADHD-C) (Association 2013). Despite its high prevalence, the precise etiology and pathogenesis of ADHD remains unclear.

To study ADHD, Tan L et al. designed a novel network based method. In this method, rsfMRI images from ADHD200-NYU dataset was firstly segmented into 351 regions according to CC400 atlas. CC400 atlas was generated via a two-level spatially constrained spectral clustering algorithm using the ADHD-200 dataset (Dai et al. 2012; Colby et al. 2012; Sato et al. 2012a, b; 2013), and was made publicly available by NeuroBureau at the competition website. Secondly, functional volume, a feature extraction method invited by the author (Tan 2015), was used to extract features from each region. Functional volume counts non-zero fALFF coefficients voxels as a measure of brain volume that was actually active during fMRI imaging. Thirdly, the region mean functional volume was calculated across all brain images. Lastly, Pair-wised Pearson's correlation between the mean functional volume was calculated to construct a 351×351 connectivity matrix. The matrix was transformed to a binary graph (Fig. 13.6), where nodes represent brain regions and edges represent undirected connections.

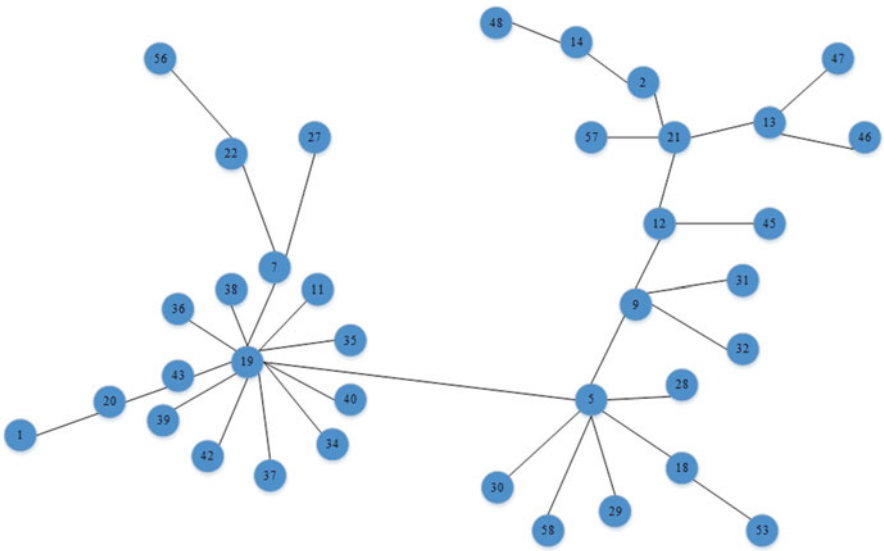


Fig. 13.6 An example of a functional brain network generated from fMRI data. The whole brain is parcellated into 116 brain regions based on AAL template

For the topological analysis, global efficiencies and local efficiencies of brain networks are compared with those of the random networks (i.e., nodes are randomly connected) and regular networks (i.e., nodes are connected to a fixed number of other nodes). The patterns of the changes in the network efficiency supported the notion that ADHD promoted a shift from small-world networks toward regular networks. Despite several recent successes, there are still many challenges on applying network approaches to study brain diseases. Further studies are necessary to address questions such as why there is a network topological alteration, when the changes in network topology first occur.

13.4.4 Network Analysis in Electronic Health Record

Network analysis has been applied on electronic health record (EHR) study. As early as 1863, clinical data has been believed to be capable of revealing more of the relative value of particular operations and models of treatment (Nightingale 1863). However, traditional paper-based health record apparently inhibits effective usage of those data, considering the difficulties of data collection and analysis. Nowadays, an EHR is a systematized collection of patient and population's health information, which are electronically stored in a digital format (Gunter and Terry 2005). With an EHR system, data would be collected quickly and are constantly being accumulated, making it possible to reuse those valuable data with a goal of improving health care quality.

EHR system commonly treats one patient as an entity, which is associated with a range of data, including demographics, medical history, allergies, immunization status, laboratory test results, radiology images, etc (Dick et al. 1997). These records can be shared among health care providers, insurance companies, and pharmacy. Utilizing Internet, information in EHR can be exchanged timely and accurately. Secondary or research use of EHR data has been making rapid progress in recent years. Research on EHR data can be generally categorized into clinical and translational research, public health surveillance for emerging threats, and healthcare quality measurement and improvement (Safran et al. 2007). It is clear that EHR data may not directly lead to knowledge. Various data mining approaches are necessary to extract knowledge from those large of amount of data. Network analysis is one of analysis approaches to reveal associations, patterns and trends hidden in the EHR data.

One of EHR applications is diagnose association network (Hanauer et al. 2009). In this work, 1.5 million free text problem summary list diagnoses of about 327,000 patients were obtained from an EHR system. With text analysis, 20,705 unique diagnose terms were selected based on all summary notes. Due to the variability of wording in the free text summary, manual text processing was further applied to combine same concept with different wording. These data were loaded into their Molecular Concept Map (MCM) application. Odds ratios (ORs) and p-value were calculated for all pairwise associations. Figure 13.7 displays a diagnosis association network with problem category.

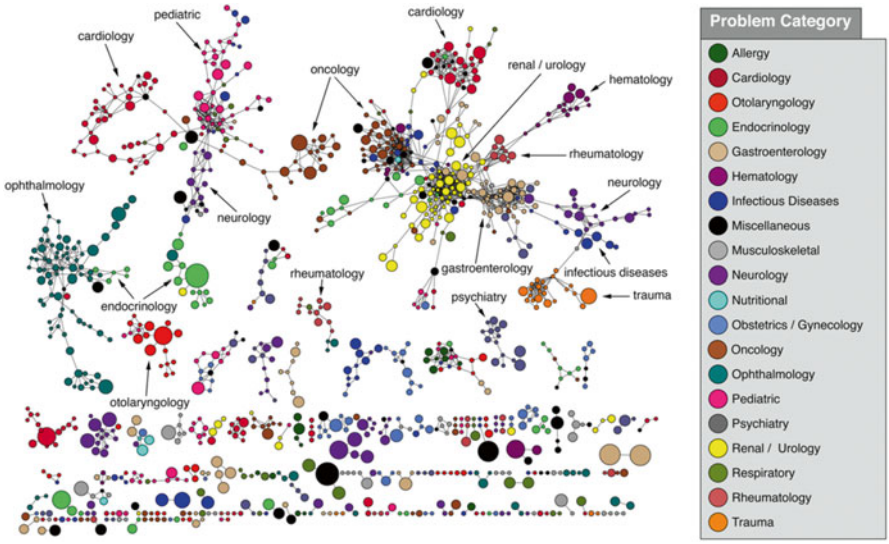


Fig. 13.7 Diagnosis association network with 1106 nodes and 1939 associations (ORs > 100.0 and p-value $<1.0\times10^{-10}$). Size of nodes reflects number of times it appears in the summary lists, and colored is marked according to category. This figure was originally published in (Hanauer et al. 2009), and is reused with permission

Constructed diagnosis association network consists of many well-known associations, as well as unexpected associations. While those known associations could be good validation tools to support the significance of the network, unexpected associations may provide new knowledge to us. Figure 13.8 demonstrates two exemplary sub-network of the overall diagnosis association network. Figure 13.8a shows a graph with centered diagnosis term "non-insulin dependent diabetes mellitus". All the associated diagnosis terms, such as "hypertension", "hyperlipidemia", and "coronary artery disease" are the most common terms in the problem summary lists related to diabetes mellitus. Figure 13.8b shows a sub-network with unexpected associations. This graph includes a few known terms related to gynecology. However, "irritable bowel" and "fibromyalgia" are not expected. Intriguingly, recent independent reports (Arnold et al. 2006, 2007) demonstrated the association between "vulvodynia" and both "Fibromyalgia" and "Irritable bowel". This indicates the association network may be used to predict new diagnosis terms association that yet to be well-known so far.

Even though network-analysis has been successfully applied in EHR data mining, there are still many challenges in the secondary use of EHR. For example, quality and completeness of data are often problematic in the data. Also, free text diagnosis reports bring difficulties in effective text mining. How to prevent patient privacy is another issue often discussed. Further studies are needed to address both technical and clinical questions.

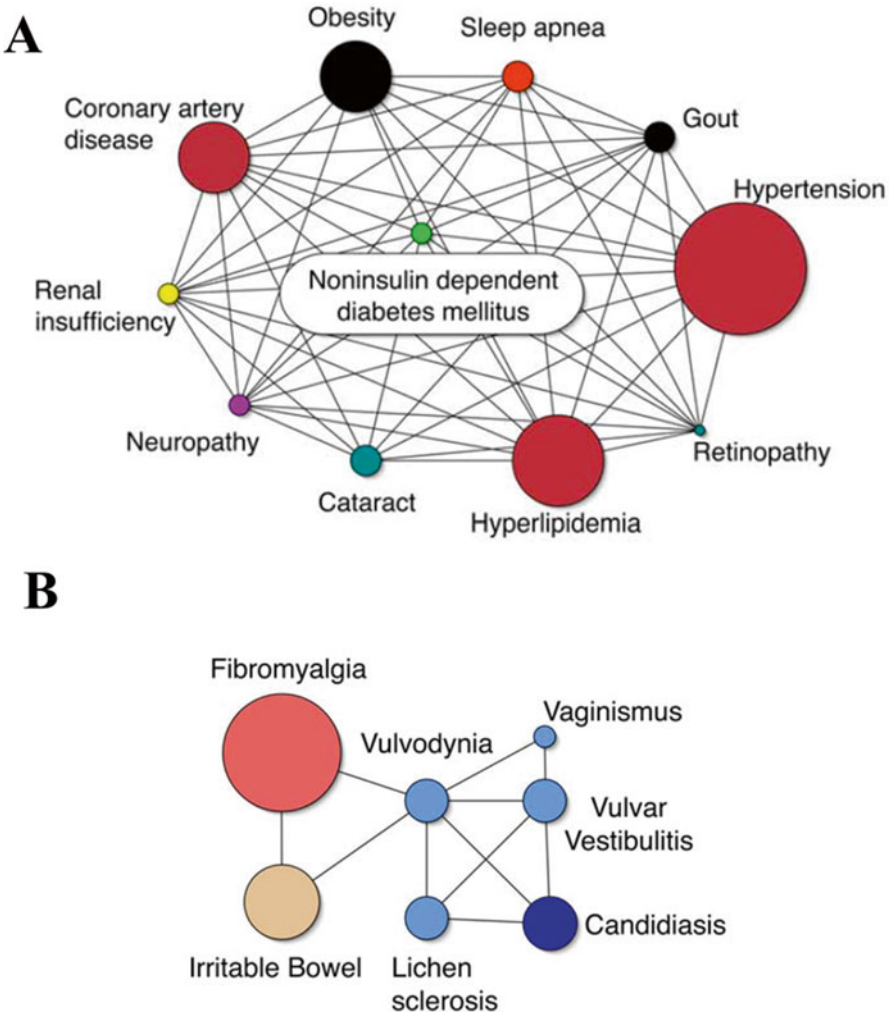


Fig. 13.8 Examples of diagnosis associations. (**a**) Sub-network with well-known clinical associations. (**b**) Sub-network with unexpected associations. Figures were obtained from (Hanauer et al. 2009) with permission

13.5 Summary

In this chapter, we discussed general biological network analysis and the applications of analyzing network topology and subnetwork structures. We also provided four frontier examples of network analysis in understanding pediatric diseases. In general, biological network analysis has been proven useful in applications such as protein function prediction, guiding wet-lab experiment design, and disease status and prognostic biomarker discovery. And the applications of network analysis have

gone beyond molecular entities to characterize many other complex relationships between entities, e.g., brain regions, in biomedical research. Applying network analysis in understanding and treating complex diseases such as cancer has just commenced in recent years. General applications of network analysis in pediatric diseases are sporadic and have not yet been systematically investigated. Much more efforts are in need in the field, and we foresee major breakthroughs using knowledge derived from network analysis and applications for pediatric disease diagnosis and treatment.

References

Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. Nature. 2000;406:378–82.

Arnold LD, Bachmann GA, Kelly S, Rosen R, Rhoads GG. Vulvodynia: characteristics and associations with co-morbidities and quality of life. Obstetrics and Gynecology. 2006;107:617.

Arnold LD, Bachmann GA, Rosen R, Rhoads GG. Assessment of vulvodynia symptoms in a sample of US women: a prevalence survey with a nested case control study. Am J Obstet Gynecol. 2007;196: 28. e1-28. e6.

Association, American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub, 2013.

Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.

Barter PJ, Nicholls S, Rye KA, Anantharamaiah GM, Navab M, Fogelman AM. Antiinflammatory properties of HDL. Circ Res. 2004;95:764–72.

Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In International AAAI conference on Weblogs and Social Media. 2009.

Batagelj V, Mrvar A. Pajek – analysis and visualization of large networks. Graph Drawing Software. 2004:77–103.

Berardini TZ, Khodiyar VK, Lovering RC, Talmud P. The gene ontology in 2010: extensions and refinements. Nucleic Acids Res. 2010;38:D331–5.

Boden WE. High-density lipoprotein cholesterol as an independent risk factor in cardiovascular disease: assessing the data from Framingham to the Veterans Affairs High – Density Lipoprotein Intervention Trial. Am J Cardiol. 2000;86:19L–22L.

Brun C, Herrmann C, Guenoche A. Clustering proteins from interaction networks for the prediction of cellular functions. BMC Bioinformatics. 2004;5:95.

Chen C-A, Chung W-C, Chiou Y-Y, Yang Y-J, Lin Y-C, Ochs HD, Shieh CC. Quantitative analysis of tissue inflammation and responses to treatment in immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome, and review of literature. J Microbiol, Immunol Infect. 2015.

Colby JB, Rudie JD, Brown JA, Douglas PK, Cohen MS, Shehzad Z. Insights into multimodal imaging classification of ADHD. Front Syst Neurosci. 2012;6:59.

Corominas R, Yang X, Lin GN, Kang S, Shen Y, Ghamsari L, Broly M, Rodriguez M, Tam S, Trigg SA, Fan C, Yi S, Tasan M, Lemmens I, Kuang X, Zhao N, Malhotra D, Michaelson JJ, Vacic V, Calderwood MA, Roth FP, Tavernier J, Horvath S, Salehi-Ashtiani K, Korkin D, Sebat J, Hill DE, Hao T, Vidal M, Iakoucheva LM. Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. Nat Commun. 2014;5:3650.

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kataskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Research. 2010;39:691–97.

Cuchel M, Rader DJ. Macrophage reverse cholesterol transport key to the regression of atherosclerosis? Circulation. 2006;113:2548–55.

Dai D, Wang J, Hua J, He H. Classification of ADHD children through multimodal magnetic resonance imaging. Front Syst Neurosci. 2012;6:63.

Davidson WS, Gangani RA, Silva D, Chantepie S, Lagor WR, Chapman MJ, Kontush A. Proteomic analysis of defined HDL subpopulations reveals particle-specific protein clusters relevance to antioxidative function. Arteriosclerosis, Thrombosis, and Vascular Biology. 2009;29:870–76.

Dick RS, Steen EB, Detmer DE. The computer-based patient record:: an essential technology for health care. Washington, DC: National Academies Press; 1997.

Emig D, Salomonis N, Baumbach J, Lengauer T, Conklin BR, Albrecht M. AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. Nucleic Acids Research. 2010;38:W755–W62.

Franceschini G, Maderna P, Sirtori CR. Reverse cholesterol transport: physiology and pharmacology. Atherosclerosis. 1991;88:99–107.

Gan Z, Wang J, Salomonis N, Stowe JC, Haddad GG, McCulloch AD, Altintas I, Zambon AC. MAAMD: a workflow to standardize meta-analyses and comparison of affymetrix microarray data. BMC Bioinformatics. 2014;15:1–11.

Girvan M, Newman ME. Community structure in social and biological networks. Proc Natl Acad Sci U S A. 2002;99:7821–6.

Goel R, Harsha HC, Pandey A, Prasad TS. Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. Mol Biosyst. 2012;8:453–63.

Gordon S, Durairaj A, Jason LL, Sean Davidson W. High-density lipoprotein proteomics: identifying new drug targets and biomarkers by understanding functionality. Current Cardiovascular Risk Reports. 2010a;4:1–8.

Gordon SM, Deng J, Jason Lu L, Sean Davidson W. Proteomic characterization of human plasma high density lipoprotein fractionated by gel filtration chromatography. Journal of Proteome Research. 2010b;9:5239–49.

Gordon SM, Li H, Zhu X, Shah AS, Lu LJ, Sean Davidson W. A comparison of the mouse and human lipoproteome: suitability of the mouse model for studies of human lipoproteins. Journal of Proteome Research. 2015;14:2686–95.

Guelzim N, Bottani S, Bourgine P, Kepes F. Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet. 2002;31:60–3.

Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. Journal of Medical Internet Research. 2005;7:e3.

Hanauer DA, Rhodes DR, Chinnaiyan AM. Exploring clinical associations using '-omics' based enrichment analyses. PLoS One. 2009;4:e5203.

Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402:C47–52.

Heller M, Stalder D, Schlappritzi E, Hayn G, Matter U, Haeberli A. Mass spectrometry – based analytical tools for the molecular protein characterization of human plasma lipoproteins. Proteomics. 2005;5:2619–30.

Homsy J, Zaidi S, Shen Y, Ware JS, Samocha KE, Karczewski KJ, DePalma SR, McKean D, Wakimoto H, Gorham J, Jin SC, Deanfield J, Giardini A, Porter GA, Kim R, Bilguvar K, López-Giráldez F, Tikhonova I, Mane S, Romano-Adesman A, Qi H, Vardarajan B, Ma L, Daly M, Roberts AE, Russell MW, Mital S, Newburger JW, William Gaynor J, Breitbart RE, Iossifov I, Ronemus M, Sanders SJ, Kaltman JR, Seidman JG, Brueckner M, Gelb BD, Goldmuntz E, Lifton RP, Seidman CE, Chung WK. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science. 2015;350:1262–66.

Hu Z, Hung JH, Wang Y, Chang YC, Huang CL, Huyck M, DeLisi C. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res. 2009;37:W115–21.

Hull J, Campino S, Rowlands K, Chan M-S, Copley RR, Taylor MS, Rockett K, Elvidge G, Keating B, Knight J, Kwiatkowski D. Identification of common genetic variation that modulates alternative splicing. PLoS Genet. 2007;3:e99.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001;411:41–2.

Kao HL, Gunsalus KC. Browsing multidimensional molecular networks with the generic network browser (N-Browse). Curr Protoc Bioinformatics, Chapter 9: Unit 9 11. 2008.

Karlsson H, Leanderson P, Tagesson C, Lindahl M. Lipoproteomics II: Mapping of proteins in high – density lipoprotein using two – dimensional gel electrophoresis and mass spectrometry. Proteomics. 2005;5:1431–45.

Kerstjens-Frederikse WS, van de Laar IMBH, Vos YJ, Verhagen JMA, Berger RMF, Lichtenbelt KD,Wassink-Ruiter JSK, van der Zwaag PA, du Marchie Sarvaas GJ, Bergman KA, Bilardo CM, Roos-Hesselink JW, Janssen JHP, Frohn-Mulder IM, van Spaendonck-Zwarts KY, van Melle JP, Hofstra RMW, Wessels MW. Cardiovascular malformations caused by NOTCH1 mutations do not keep left: data on 428 probands with left-sided CHD and their families. Genet Med. 2016.

King AD, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering. Bioinformatics. 2004;20:3013–20.

Koh K-N, Im HJ, Chung N-G, Cho B, Kang HJ, Shin HY, Lyu CJ, Yoo KH, Koo HH, Kim H-J, Baek HJ, Kook H, Yoon HS, Lim YT, Kim HS, Ryu KH, Seo JJ, Party the Korea Histiocytosis Working. Clinical features, genetics, and outcome of pediatric patients with hemophagocytic lymphohistiocytosis in Korea: report of a nationwide survey from Korea Histiocytosis Working Party. European Journal of Haematology. 2015;94:51–9.

Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, Mélius J, Waagmeester A, Sinha SR, Miller R, Coort SL, Cirillo E, Smeets B, Evelo CT, Pico AR. WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Research. 2016;44:D488–D94.

Lewis GF, Rader DJ. New insights into the regulation of HDL metabolism and reverse cholesterol transport. Circulation research. 2005;96:1221–32.

Li H, Gordon SM, Zhu X, Deng J, Swertfeger DK, Davidson WS, Lu LJ. Network-based analysis on orthogonal separation of human plasma uncovers distinct high density lipoprotein complexes. J Proteome Res. 2015;14:3082–94.

Lu CX, Gong HR, Liu XY, Wang J, Zhao CM, Huang RT, Xue S, Yang YQ. A novel HAND2 loss-of-function mutation responsible for tetralogy of Fallot. International Journal of Molecular Medicine. 2016;37:445–51.

Lucas CL, Yu Z, Venida A, Wang Y, Hughes J, McElwee J, Butrick M, Matthews H, Price S, Biancalana M, Wang X, Richards M, Pozos T, Barlan I, Ahmet O, Koneti Rao V, Su HC, Lenardo MJ. Heterozygous splice mutation in PIK3R1 causes human immunodeficiency with lymphoproliferation due to dominant activation of PI3K. The Journal of Experimental Medicine. 2014;211:2537–47.

Maslov S, Sneppen K. Specificity and stability in topology of protein networks. Science. 2002;296:910–3.

Mineo C, Deguchi H, Griffin JH, Shaul PW. Endothelial and antithrombotic actions of HDL. Circulation Research. 2006;98:1352–64.

Naqvi TZ, Shah PK, Ivey PA, Molloy MD, Thomas AM, Panicker S, Ahmed A, Cercek B, Kaul S. Evidence that high-density lipoprotein cholesterol is an independent predictor of acute platelet-dependent thrombus formation. The American Journal of Cardiology. 1999;84:1011–17.

Negre-Salvayre A, Dousset N, Ferretti G, Bacchetti T, Curatola G, Salvayre R. Antioxidant and cytoprotective properties of high-density lipoproteins in vascular cells. Free Radical Biology and Medicine. 2006;41:1031–40.

Nightingale F. Notes on hospitals (Longman, Green, Longman, Roberts, and Green). 1863.

Nofer J-R, Kehrel B, Fobker M, Levkau B, Assmann G, von Eckardstein A. HDL and arteriosclerosis: beyond reverse cholesterol transport. Atherosclerosis. 2002;161:1–16.

O'Madadhain J, Fisher D, White S, Boey YB. The JUNG (Java Universal Network/Graph) framework. In: Technical report UCI-ICS 03–17. Irvine: UC Irvine; 2003. p. 03–17.

Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature. 2005;435:814–8.

Ravasz E. Detecting hierarchical modularity in biological networks. Methods Mol Biol. 2009;541:145–60.

Rezaee F, Bruno C, Han J, Levels M, Speijer D, Meijers J. Proteomic analysis of high-density lipoprotein. Proteomics. 2006;6:721–30.

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005;437:1173–8.

Safran C, Meryl B, Edward Hammond W, Labkoff S, Markel-Fox S, Tang PC, Detmer DE. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association. 2007;14:1–9.

Sato JR, Hoexter MQ, Castellanos XF, Rohde LA. Abnormal brain connectivity patterns in adults with ADHD: a coherence study. PLoS One. 2012a;7:e45671.

Sato JR, Hoexter MQ, Fujita A, Rohde LA. Evaluation of pattern recognition and feature extraction methods in ADHD prediction. Front Syst Neurosci. 2012b;6:68.

Sato JR, Takahashi DY, Hoexter MQ, Massirer KB, Fujita A. Measuring network's entropy in ADHD: a new approach to investigate neuropsychiatric disorders. Neuroimage. 2013;77:44–51.

Schaffer AE, Eggens VRC, Caglayan AO, Reuter MS, Scott E, Coufal NG, Silhavy JL, Xue Y, Kayserili H, Yasuno K, Rosti RO, Abdellateef M, Caglar C, Kasher PR, Cazemier JL, Weterman MA, Cantagrel V, Cai N, Zweier C, Altunoglu U, Bilge Satkin N, Aktar F, Tuysuz B, Yalcinkaya C, Caksen H, Bilguvar K, Fu X-D, Trotta C, Gabriel S, Reis A, Gunel M, Baas F, Gleeson JG. CLP1 founder mutation links tRNA splicing and maturation to cerebellar development and neurodegeneration. Cell. 2014;157:651–63.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

Soreq L, Guffanti A, Salomonis N, Simchovitz A, Israel Z, Bergman H, Soreq H. Long non-coding RNA and alternative splicing modulations in Parkinson's leukocytes identified by RNA sequencing. PLoS Comput Biol. 2014;10:e1003517.

Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A. 2003;100:12123–8.

Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003;302:249–55.

Tan L. Identification of disease biomarkers from brain fMRI data using machine learning techniques: applications in sensorineural hearing loss and attention deficit hyperactivity disorder. University of Cincinnati. 2015.

Vaisar T, Pennathur S, Green PS, Gharib SA, Hoofnagle AN, Cheung MC, Byun J, Vuletic S, Kassim S, Singh P. Shotgun proteomics implicates protease inhibition and complement activation in the antiinflammatory properties of HDL. The Journal of Clinical Investigation. 2007;117:746–56.

Wagner A, Fell DA. The small world inside large metabolic networks. Proc Biol Sci. 2001;268:1803–10.

Wang G-S, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet. 2007;8:749–61.

Watson AD, Berliner JA, Hama SY, La Du BN, Faull KF, Fogelman AM, Navab MOHAMAD. Protective effect of high density lipoprotein associated paraoxonase. Inhibition

of the biological activity of minimally oxidized low density lipoprotein. Journal of Clinical Investigation. 1995;96:2882.

Wu J, Vallenius T, Ovaska K, Westermarck J, Makela TP, Hautaniemi S. Integrated network analysis platform for protein-protein interactions. Nat Methods. 2009;6:75–7.

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, Krainer AR, Jojic N, Scherer SW, Blencowe BJ, Frey BJ. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347:1254806.

Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, Sun S, Yang F, Shen YA, Murray RR, Spirohn K, Begg BE, Duran-Frigola M, MacWilliams A, Pevzner SJ, Zhong Q, Trigg SA, Tam S, Ghamsari L, Sahni N, Yi S, Rodriguez MD, Balcha D, Tan G, Costanzo M, Andrews B, Boone C, Zhou XJ, Salehi-Ashtiani K, Charloteaux B, Chen AA, Calderwood MA, Aloy P, Roth FP, Hill DE, Iakoucheva LM, Xia Y, Vidal M. Widespread expansion of protein interaction capabilities by alternative splicing. Cell. 2016;164:805–17.

Yip KY, Yu H, Kim PM, Schultz M, Gerstein M. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. Bioinformatics. 2006;22:2968–70.

Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol. 2007;3:e59.

Zhang M, Deng J, Fang C, Zhang X, Lu LJ. Molecular network analysis and applications. In: Alterovitz G, Ramoni M, editors. Knowledge-based bioinformatics: from analysis to interpretation. Chichester: Wiley; 2010.

Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, Johnson AD, Levy D, O'Donnell CJ. Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat Genet. 2015;47:345–52.

Part III
Genomic Applications

Chapter 14
Genetic Technologies and Causal Variant Discovery

Phillip J. Dexheimer, Kenneth M. Kaufman, and Matthew T. Weirauch

Abstract The widespread availability of next-generation sequencing (NGS) has transformed our understanding of human genetic variation and its impact on human health. This chapter describes the most common DNA sequencing technologies available to research and clinical laboratories today, and resources for interpreting the functional impact of genetic variants identified with these technologies. Targeted genetic capture techniques were developed to dramatically decrease the cost of determining variant genotypes, although as prices decrease many of the advantages of targeted experiments diminish. Targeted whole exome sequencing is used to identify variants that alter the amino acid sequence of a protein. Sequencing can also be performed for the whole genome, enabling the identification of variants that fall within non-coding regions, which might alter the expression of a gene by disrupting regulatory sequences such as transcription factor binding sites. Careful consideration of the study design, particularly the use of prior knowledge about the phenotype in question, family history, and the availability of affected and unaffected family members, increases the chances that meaningful results are obtained. Identifying variants in the short, error-prone sequencing reads generated by modern technologies is challenging, although software packages exist that mitigate the most common types of errors. The most difficult task in analyzing results is interpreting the functional impact of putative variants and differentiating between clinically reportable variants and variants of unknown significance (VUS) or variants within genes of unknown significance (GUS). Additional information, such as population allele frequencies, genetic inheritance patterns, and functional genomics data, can help to identify the variants most likely to be involved in disease pathogenesis.

P.J. Dexheimer (✉)
Division of Biomedical Informatics, Cincinnati Children's Hospital,
3333 Burnet Avenue, Cincinnati, OH 45229, USA
e-mail: phillip.dexheimer@cchmc.org

K.M. Kaufman, Ph.D. • M.T. Weirauch, Ph.D.
Center for Autoimmune Genomics and Etiology (CAGE)
and Division of Biomedical Informatics, Cincinnati Children's Hospital,
3333 Burnet Avenue, Cincinnati, OH 45229, USA
e-mail: kenneth.kaufman@cchmc.org; matthew.weirauch@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_14

Keywords Alignment • DNA sequencing • Exome • Genome • Variant detection

14.1 Introduction

Sequencing technologies have rapidly evolved over the last two decades. Traditional chain-termination, or "Sanger" sequencing, continues to be an important part of both clinical and research applications. Sanger sequencing has a very low error rate, and is considered the "gold standard" in identifying variants. However, per-base costs remain very high and throughput remains very low compared to more recent technologies. Modern applications typically use one of the many "next-generation" sequencing (NGS) technologies, which are characterized by extremely high throughput, low per-base costs, and a higher error rate.

The widespread availability of these technologies has transformed our understanding of human genetic variation and its impact on human health. In this chapter, we describe several of the most common technologies available to research and clinical laboratories today, considerations for effective study design, and resources for interpreting the functional impact of variants identified with these technologies.

14.2 Overview of Next-Generation Sequencing Technologies

A variety of alternative technologies have come to market since ~2005. Despite the immense variety they represent, these technologies are typically lumped together under the umbrella of "next-generation". In this section, we provide an overview of several of these disparate technologies, focusing on those marketed under the brands Illumina, Ion Torrent, and Pacific Biosciences (Table 14.1). This section is not intended to be an exhaustive survey of the field, and several interesting technologies are left out. Rather, our intent is to illustrate the variety of techniques used to achieve large scale sequencing and discuss some of the most widespread machines.

14.2.1 Sequencing by Synthesis (Illumina)

Sequencing by synthesis was developed by Solexa, Ltd, which was acquired by Illumina in 2007. For this reason, the terms "sequencing by synthesis", "Illumina sequencing", and "Solexa sequencing" are used interchangeably. The fundamental idea behind this technology is to measure the order of nucleotides incorporated during second strand synthesis (Bentley et al. 2008).

Table 14.1 Comparison of popular next-generation sequencing instruments

Instrument	Technology	Read length	Approximate cost/Gb	Instrument list cost
Illumina MiSeq	Sequencing by synthesis	Up to 300 bp	$90	$125,000
Illumina NextSeq	Sequencing by synthesis	Up to 150 bp	$30	$250,000
Illumina HiSeq 2500	Sequencing by synthesis	Up to 150 bp	$30–45	$740,000
Illumina HiSeq X Ten	Sequencing by synthesis	Up to 150 bp	$7	$1,000,000
Ion PGM	Semiconductor sequencing	Average 400 bp	$375	$50,000
Ion Proton	Semiconductor sequencing	Average 200 bp	$120	$150,000
Pacific Biosciences Sequel	SMRT sequencing	Average 10–15 kb	$85	$350,000

Standard library preparation begins by randomly fragmenting DNA to a size of several hundred nucleotides. Custom adapters that contain primer sites for amplification and sequencing reactions are ligated to the resulting double-stranded fragments. These adapters are complementary to each other on the end closest to the fragment of interest, but are intentionally non-complementary for much of their length. These "floppy ends" result in a Y-shaped adapter, which is critical for the efficiency of library preparation. More recently, custom transposons have been used to perform the fragmentation and ligation steps in a single tagmentation reaction.

The sequencing substrate, called a flowcell, is a glass slide covered in a field of fixed oligonucleotides. The single-stranded library is introduced across the flowcell at a rate and concentration such that individual molecules hybridize randomly across the entire field. The flowcell then undergoes repeated rounds of bridge amplification. In this process, individual molecules are bent over to allow their free end to hybridize to a different probe on the surface of the flowcell, followed by second strand synthesis and denaturing. After several rounds of this process, the individual molecule is replicated tens of thousands of times in a very localized area of the flowcell, creating distinct clusters of unique fragments.

Once these clusters have been formed, sequencing can begin. Modified nucleotides that include a fluorescent marker and a blocking element on the 3′ end are introduced, which are incorporated into second strand synthesis. The presence of the blocker guarantees that only a single nucleotide can be incorporated onto any fragment, permitting the mixture to include all four nucleotides. Once incorporation is completed, the fluorescent markers are excited and imaged. A chemical is introduced that cleaves the blocker from the partial second strand, and the process begins again with a new flow of nucleotides. Each cycle thus produces an image of the flowcell with colored dots at the position of each cluster. Critically, the clusters

themselves do not move between rounds. By registering and analyzing a series of these images, the sequence of each distinct cluster can be determined.

This technique has been used, with minor modifications, in all of Illumina's sequencers. Improvements have focused on increasing cluster density without making adjacent clusters indistinguishable and reducing the time necessary for imaging. The most recent iterations use a patterned flowcell with preconfigured clusters inside nanowells, which improves both density and separation of clusters. Nearly all errors in this technology affect only a single nucleotide; erroneous insertions and deletions are rare. Because of the extended time spent in synthesis, error rates characteristically increase over time as the polymerase begins to fail.

14.2.2 Semiconductor-Based Sequencing (Ion Torrent)

A very different sequencing technique was developed by Ion Torrent Systems, Inc. Rather than relying on imaging or reporter assays, Ion developed a semiconductor sensor that detects ions liberated during a standard nucleotide incorporation event – thus yielding a much simpler, cheaper system (Rothberg et al. 2011). After a series of acquisitions and mergers, Ion Torrent is currently a part of Thermo Fisher.

Library preparation proceeds similarly to the Illumina system, with fragmentation, ligation, and amplification steps. However, the amplification substrate is a magnetic bead rather than a glass slide. Amplification is performed in an emulsion Polymerase Chain Reaction (emPCR) consisting of microdroplets containing an adaptor-ligated library and beads within oil. Concentrations and both fragments and library must be carefully adjusted to ensure that each droplet contains no more than one molecule of template and one bead. Following amplification, the emulsion is broken and amplified beads are separated from empty beads. Each bead contains approximately 800,000 clonal copies of template.

These beads are then loaded onto a disposable semiconductor containing millions of microwells, sized such that only a single bead fits into each well. The actual sequencing reaction is very similar to pyrosequencing. Individual nucleotides are introduced and incorporated onto the second strand by a bound DNA polymerase. As part of the incorporation reaction, a single proton is liberated per nucleotide. This proton changes the pH in the well by a minute amount, which is detected by a sensor in the bottom of the well.

The advantages of this technique are largely in its simplicity. The sequencer requires no optics, no specialized reagents, no extended time to image. The sensor uses standard semiconductor construction techniques, which have been optimized and refined over the last several decades. The result is a relatively inexpensive sequencer that runs very quickly. However, like pyrosequencing, the Ion sequencers are easily confused by homopolymer runs in the template. In principle, the change in pH is directly correlated to the number of incorporated nucleotides in any one event, but in practice it is difficult to reliably determine homopolymer lengths over about five nucleotides.

14.2.3 SMRT Sequencing (Pacific Biosciences)

Single Molecule Real Time (SMRT) Sequencing is a more recent technique, developed by Pacific Biosciences (Eid et al. 2009). As the name implies, this technique does not rely on the clonal amplification of a sequencing library; instead, an individual molecule of DNA is directly sequenced. Read lengths are also vastly improved, in the tens of kilobases range as opposed to the hundreds of bases produced by other technologies.

Library preparation is again similar in spirit to the other technologies, though instead of ligating sequences for amplification or hybridization, a closed loop is ligated onto each end of the double-stranded library, producing a "dumbbell" shape and effectively circularizing the template.

The sequencing substrate is a specially engineered surface containing an array of zero-mode waveguide (ZMW) nanowells. These wells are small enough that they interfere with light, allowing only the bottom 20–30 nm to be illuminated from below. A single DNA polymerase is bound to each template molecule, and then attached to the bottom of a ZMW. Nucleotides with phospholinked fluorescent tags are then introduced and allowed to incorporate. Each incorporation event causes the release of the associated fluorescent tag within the viewable region of the ZMW, which is imaged in real time.

Because the polymerase is not repeatedly interrupted, very long reads are possible. Further, because the template is circularized, each molecule can be sequenced several times. These factors help to mitigate the relatively high error rate of SMRT sequencing, by permitting consensus sequencing of single molecules.

14.3 Targeted vs Whole Genome Sequencing

Several approaches can be used for leveraging NGS technology to detect disease causing variants or explore various biological processes. The most comprehensive approach is to perform whole genome sequencing (i.e., sequencing the entire genome of an individual). An alternative approach is to use DNA probes to isolate and enrich only certain regions of the genome of interest, which are then sequenced. For example, this approach is widely used to target the exons of all known genes in whole exome sequencing. Alternatively, a custom targeted array may be used to enrich for a small number of genes of interest. Whole genome sequencing generates the greatest amount of information about the genome. These data sets offer the potential to not only identify variants that alter the amino acid sequence of a protein but also variants that could alter gene regulatory regions. In addition, whole genome sequence can be used to identify chromosomal rearrangements and copy number variations. While complete coverage of the genome might be beneficial in some situations, there are some drawbacks. The cost of whole genome sequencing is higher than a more targeted approach such as whole exome sequencing. The amount

of data generated is ~50 times greater than whole exome sequencing, which results in greater bioinformatic burden. Interpretation of the functional impact of non-coding variants in introns or intergenic regions is much more difficult than amino acid-altering variants (discussed below). Often, the overall coverage (i.e., average number of sequencing reads per base) is lower in whole genome data sets compared to targeted sequencing approaches. This lower coverage can result in less reliable variant calling and increase the number of false positive results.

Whole exome sequencing is often used due to the lower cost, increased coverage and easier interpretation of the results. Often, the first variants to be analyzed are those that alter the amino acid sequence of a protein. Due to our understanding of gene structure and codon usage, it is much easier to identify variants that could have a biological impact on protein function by altering amino acids. This approach will be successful if the disease relevant variant is located in a targeted exon. However, if the disease relevant variant is located in an intron or gene regulatory region it will not be detected, as only ~5 % of the genome is sequenced in whole exome sequencing.

A viable approach to address the shortcomings of both whole genome and whole exome sequencing is to utilize a custom targeted sequencing approach. In such techniques, if a small region of the genome is suspected to contain relevant variants, only this region will be enriched and sequenced. A similar strategy can be used if a list of candidate genes are identified that are thought to be involved in disease. In custom targeted sequencing approaches, introns, exons and intergenic regions can be targeted. Often these data sets are much smaller than whole exome sequencing but provide even more coverage than exome sequencing. This results in overall lower cost and greater accuracy in variant calls.

14.4 Alignment and Variant Calling

A variety of current and historical market pressures mean that, as of this writing, short-read sequencers, primarily from Illumina, dominate both the research and clinical landscapes. The individual sequences produced by these sequencers are too short to reliably assemble into a complete genome (Alkan et al. 2011), so current variant calling techniques require that sequences first be aligned to a reference genome. We note that the increasing availability of long-read sequencers makes hybrid assembly/alignment techniques more feasible (Chaisson et al. 2015), but we do not address those techniques here.

The volume of short reads generated by a single experiment demands that alignment algorithms execute incredibly quickly, and the older heuristic-based algorithms (Altschul et al. 1990; Kent 2002) lack sufficient speed and sensitivity in such a setting. The field has coalesced around several programs that use the Burrows-Wheeler Transform to rapidly find exact matches. Successful programs (Langmead et al. 2009; H. Li and Durbin 2009) also include heuristic-based methods to permit inexact matches (within user-specified bounds). A key output of these algorithms is

a measure of confidence in the mapping, quantified as a mapping quality. This metric is primarily used to distinguish reads that have an ambiguous origin – due to repetitive sequence in the genome, for instance – from those which can confidently be assigned to a single position in the genome.

Identifying variation in a sequenced patient is a difficult balance between sensitivity and specificity. Biases and errors introduced during sequencing and alignment can lead to artifactual variants or mask true variants, misleading naïve algorithms. The most successful algorithms use a variety of techniques to mitigate known biases and meaningfully balance sensitivity and specificity.

One of the most popular variant calling algorithms is employed in the Genome Analysis Toolkit (GATK) (DePristo et al. 2011; 1000 Genomes Project et al. 2012). The techniques used in GATK are illustrative of the best practices in the field. Before variant calling even begins, the GATK pipeline includes several steps to minimize known biases and errors introduced by the sequencer and short read aligners. The variant calling process itself is designed to be as sensitive as possible; specificity is managed by a *post hoc* filtering process.

14.4.1 Variant Calling Data Pre-processing

Preprocessing proceeds through three phases, each addressing a specific bias. The first phase, realignment, addresses shortcomings of short read aligners in the face of insertions and deletions (indels). Aligners score gaps differently depending on which part of the sequencing read is affected, which means that multiple reads spanning a single gap will be aligned differently. During the realignment stage, putative indels are identified in the alignment. These can simply be specified as a set of known indels that have previously been identified, or based on a scan of the alignments for a particular patient searching for regions with a large amount of noise. Once the putative indels have been identified, a slower, more sensitive Smith-Waterman based aligner (Smith and Waterman 1981) can be applied to each region. Applying this correction helps to reduce the number of false positive variants in the vicinity of indels that are simply caused by misalignment.

The second phase of preprocessing deals with the presence of PCR-derived read duplicates. Several phases of library preparation include PCR amplification, particularly in a targeted sequencing experiment. If, through random chance, a particular allele is excessively amplified, it may alter the ratio of sequenced alleles enough to cause a miscall. This phase identifies putative duplicates by examining the aligned coordinates and strand of each read, and removes any such duplicates from consideration. In addition to improving the quality of eventual calls, identifying duplicates is a useful quality control metric. Libraries with an excessive amount of duplication can suffer from other unknown biases, and should be removed from further analysis.

In the final phase of preprocessing, base quality scores are adjusted. Sequencers estimate the probability that a particular base has been miscalled based on, e.g., the relative difference in fluorescent intensities. The recalibration phase attempts to empirically adjust these error rates. Base quality scores are stratified according to a set of technology-specific covariates. For Illumina sequencing, these covariates include the position of the base within the read and the local context of each base. After removing bases that align to sites of known variation, an empirical mismatch rate is derived from the bases with each combination of covariates. Comparing these empirical scores to the native quality score yields adjustments for each covariate/quality score pair. Finally, the quality score for each base is adjusted by applying every relevant covariate. In this way, the accuracy of each basecall is made substantially more accurate, making this phase of data analysis crucial for accurate variant calling.

14.4.2 Variant Calling

Variant callers typically employ Bayesian methods to account for the information present in reads while still maintaining stringency over billions of potential sites in a given experiment (R. Li et al. 2009; DePristo et al. 2011). The Bayesian use of prior probabilities allows the algorithms to properly account for the relative rarity of variation at any particular site and still call a variant when the evidence supports it. When weighing this evidence for variation, it is critical to have an accurate estimate of the quality of each base call, which is the reason for the importance of base quality recalibration.

While next-generation variant callers have typically only considered a single site at a time independently of all other sites, GATK has recently developed an approach that considers variation across a region. The 'HaplotypeCaller' module first defines regions on the order of several hundred nucleotides in length that contain significant evidence for variation. The reads in these regions are assembled into putative haplotypes, which are then aligned back against the reference genome using the Smith-Waterman algorithm in order to determine sites of variation. Each read in the region is then aligned to each potential haplotype in order to determine the likelihood of the haplotypes. From these haplotype likelihoods, a likelihood of each allele at each variant site is determined. Finally, these allele likelihoods are used as the input to a Bayesian algorithm that determines the most probable genotype of each sample at each site in the region. Since this algorithm considers all of the variation in a region at the same time, it is less likely to be misled by small structural changes that lead to misalignment and have classically caused errors in variant calling. In addition, physical phasing of variation over each considered region is output. Typically, sensitivity of calling is emphasized in this stage, and the results are filtered after the fact to enhance specificity. We discuss filtering techniques in Sect. 14.6.

14.5 Effective Study Design

Before NGS is undertaken, careful consideration needs to be taken to determine if this technology is applicable and to establish an effective study design. A typical whole exome sequencing data set for an individual will identify 50,000–70,000 variants, whereas whole genome sequencing will identify 3–4 million variants. The goal is to reduce these numbers to a few variants that either cause disease or are strong candidates to produce the phenotype being studied. If the phenotype is complex and involves common variants or environmental factors, it will be difficult to identify the causative variants due to the small sample size and large number of variants. On the other hand, if the disease is rare, there is a strong family history or an extreme phenotype, then NGS technology is likely a good approach to discover the underling genetic basis of disease.

Study design requires knowledge of the family history, availability of samples and predictions on the mode of inheritance. If the mode of inheritance follows a dominant model it will be extremely difficult to identify a causative variant due to the large number of heterozygous variants found in all samples. For dominant models, large cohorts of affected individuals and ideally controls are necessary to establish association between a specific variant and disease. Variants that follow a recessive model are more easily identified because the number of variants that fit these models are smaller. For example, only 1 out of 10,000 individuals would be expected to be homozygous for a variant with a minor allele frequency of 1 % in the general population. Such variants are easily identified and would be strong candidates for causation if the gene's function is consistent with the phenotype.

Often, a trio study design is utilized in which both parents and the proband (affected individual) are sequenced. For an effective trio study design, both parents should be unaffected with respect to the phenotype being examined. The genetic profile of the parents is used to identify variants in the proband that fit specific models of inheritance. For example, a de novo model would predict that both parents are homozygous for the reference allele and the proband is heterozygous for a polymorphism that was spontaneously generated early in development. Other models include a recessive homozygous inheritance pattern, where both parents are heterozygous for the same variant and the proband is homozygous for the alternate allele. If either parent is homozygous for a variant found in the proband, then it is discarded as a candidate variant because the parent is unaffected. A third model involves compound heterozygous variants in the same gene. Each parent contains a different heterozygous variant in the same gene. While each variant can affect gene function, the presence of a normal copy of the gene prevents disease in the parents. The proband inherits both heterozygous variants and the result is that both copies of the gene are affected, leading to disease. In a typical trio analysis, the number of candidate amino acid altering rare recessive homozygous variants or de novo variants is less than five. Compound heterozygous variants are usually found in less than 10 genes.

If parental DNA is not available for analysis, then a singleton approach can be taken. Homozygosity for rare amino acid altering variants or genes with multiple heterozygous variants can then be identified. However, the lack of parental sequence will result in lists of candidate variants that can exceed ten times the number of variants identified in a trio analysis. The ability to sequence unaffected siblings can be beneficial, since this will enable the removal of candidate variants where the same mode of inheritance can be detected in the unaffected sibling. However, often unaffected siblings do not provide enough of an advantage to warrant the cost of sequencing. More distantly related unaffected family members provide even less benefit, as the number of commonly inherited variants is less than for a sibling. Conversely, additional affected family members can be a benefit, as the more often a variant is found in multiple affected family members, the stronger the candidate is for causing disease. This is especially true the more distantly related the family members are.

14.6 Filtering Genomic Data

In order to identify disease causing variants, NGS data must be filtered to a manageable amount of variants and genes that can be evaluated and prioritized. Variants can be filtered based on a number of different characteristics. These can include quality control measurements, predicted functional effects and allele frequency based on large population genetic studies. While NGS generates highly accurate data, it must be remembered that the vast majority of data will match the reference genome and is thus discarded. The sequences of interest are the differences between the sample genome and the reference genome. Such differences can be caused by actual DNA variation or sequencing artifacts. Thus, an unfortunate consequence of calling variants is that sequencing artifacts are concentrated in the data set used for analysis.

Often the variants identified in the individuals being studied are tabulated in a VCF (variant call format) file (Danecek et al. 2011). This is a text based file that contains the chromosomal location of the variant, overall quality statistics of the variant in all samples called, as well as individual sample statistics and features of the variant. Some of the commonly reported features include the read depth (number of reads containing the variant), allele depth (number of reads with the reference and alternate allele) and a genotype quality score representing the probability that the called genotype is accurate. The VCF file format is flexible – while some features are required such as chromosome and position, other features are optional, and thus not all variant calling software will provide the same features and information. The quality measurements are often used to filter the variants and remove called variants that are likely sequencing artifacts. Higher read depth (the number of times a base was detected in the sequencing reads) results in greater sensitivity to accurately detect variants. Ideally, the average read depth should be between 75 and 100X for whole exome sequencing. Read depths below 50X increase the number of false positive results obtained during subsequent validation steps. With an average

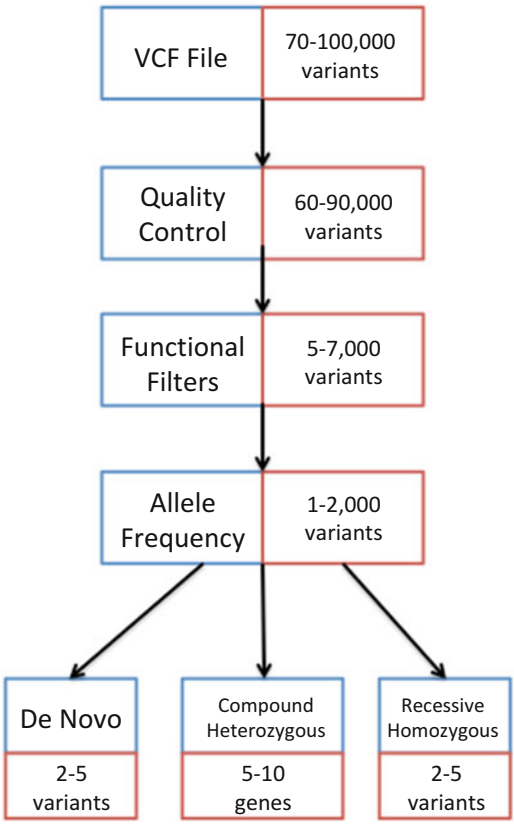


Fig. 14.1 Typical variant identification workflow - This flowchart shows a typical filtering strategy as well as representative variant counts at each stage of filtering. Refer to the text for definitions of the filters and more details

read depth >75X, a filter that requires at least 15 or 20 reads will only remove a small fraction of variant calls that are less likely to be accurate due to the low depth. Whenever data are filtered, the issue of sensitivity vs specificity arises. Any filtering introduces the chance that relevant data will be removed from the data set. On the other hand, without removing unreliable data, the rate of false positives greatly increases. A balance must be established that decreases the number of false positives and maintains the ability to detect clinically or biologically relevant variants.

A typical workflow for trio analysis of whole exome sequence data is shown in Fig. 14.1. The input is a multi-sample VCF file containing the variants detected in the parents or the proband. A multi-sample VCF file will have variant calls at all positions in which any one of the samples has a detectable variant. For trio analysis this is critically important because the proband may have a variant that is absent in one or both parents. In single sample vcf files, the parents would not have this positon called and it would be unclear if the parents are actually homozygous for this position or if the sequencing data were not adequate for variant calling. While any individual sample will have between 40,000 and 70,000 variants, a trio will often exceed 100,000 variants because it represents all variants found in any of the

samples. Usually, individual genotypes are removed based on quality control features that include a read depth <15, a genotype quality score <20 and the ratio of alternate alleles to reference alleles, in a genotype specific manner. Variants called homozygous for the alternate allele are removed if the ratio is below 85 %, heterozygous calls are removed if they are greater than 70 % or less than 30 %, and homozygous reference genotypes are removed if the ratio exceeds 15 %. Typically, these filters remove ~95 % of the sequencing artifacts at a cost of ~7–12 % of the data.

The GATK provides an alternative approach to this "hard filtering" technique, where variants must pass a series of thresholds to be considered confident. In the Variant Quality Score Recalibration (VQSR) technique, a curated set of known variants is used to distinguish biological variation from artifact. For every variant called by GATK, a set of metrics such as "average mapping quality at the site" and "strand bias" is calculated. Using the subset of known variants that are actually present in the dataset, a Gaussian mixture model is fitted to these metrics to define ranges of values that are associated with biological variation. Every called variant can then be assessed using the mixture model, yielding a log-odds score that the variant call is true. This approach has several advantages: laboratory-specific fluctuation in the metrics is properly handled, the mixture model assesses the performance of the variant across several dimensions at once rather than a single metric at a time, and the single score combined with the presence of known variants allows the analyst to directly alter the sensitivity of analysis.

The next filter often used is based on the functional characteristics of the variant. In this step, only variants that alter the amino acid sequence of a protein are retained. These include non-synonymous substitutions, stop-codon gain or loss, frame-shifts, disruption of the initiation methionine codon, or altered exon/intron splice sites. This filter removes the vast majority of variants from the data set. The next filter is based on population allele frequency data obtained from such resources as the 1000 Genomes Project (1000 Genomes Project et al. 2015), the exome sequencing project (Tennessen et al. 2012) or the ExAC database (Exome Aggregation Consortium et al. 2015). In most cases the analysis is performed under the assumption that the disease causing variant is either rare or novel. The threshold for the minor allele frequency can be based on disease prevalence. Commonly, a 1 % minor allele frequency is used as a starting point. However, even lower frequencies may be useful in identifying the best candidate disease causing variants. After these initial filtering steps, the remaining variants are screened through different genetic models of inheritance. Variants that fit the various models become candidates for causative variants.

14.7 Functional Interpretation of Variants

Variants that pass through the above filtering steps represent those most likely to be causing the disease. Interpretation of the functions of these variants is a multistep process that is extremely difficult to automate. Each variant must be evaluated based

on a number of different factors. For example, literature searches can be performed to determine if the variant or proximal gene have previously been associated with disease. This can also be achieved by interrogating relevant databases such as ClinVar (Landrum 2014). Likewise, relevance of the gene's function relevant to disease physiology, or the gene's expression levels in disease-relevant cell types should be considered. Further, it is important to consider if the variant's predicted functional effects are severe enough to be biologically relevant. For example, a premature stop codon in an early exon would most likely be more significant than an alanine to glycine substitution in the carboxyl terminus of a protein. Collectively, the answers to these and other questions can be used to determine which variant(s) might be causing disease.

Interpretation of the functional impact of variants located in non-coding regions of the genome is particularly challenging (Ward and Kellis 2012b; Paul et al. 2014; Zhang and Lupski 2015; Lee and Young 2013), which is especially pertinent given that between 85 and 93 % of disease/trait associated variants are non-coding (Hindorff et al. 2009; Maurano et al. 2012). Given their location outside of protein coding regions, such variants might act by impacting gene regulatory mechanisms. Currently, computational resources exist for examining three broad categories of regulatory mechanisms that might be affected by non-coding variants: alterations in the binding of transcription factors (TFs), RNA binding proteins (RBPs), and microRNAs (miRs). To predict the impact of non-coding variants, variants should first be partitioned into those residing within predicted mRNAs (which are most likely to affect miR and RBP binding), those located in likely promoters/enhancers (which might affect TF binding), and those unlikely to affect gene regulation (e.g., variants located within "gene deserts"). To this end, relevant functional genomics data (including RNA-seq, DNase-seq, and ChIP-seq for TFs and regulatory histone marks such as H3K27ac and H3K4me3) can be obtained from sources such as the UCSC Genome Browser (Rosenbloom et al. 2013), NIH Roadmap Epigenomics (Chadwick 2012), Cistrome (Liu et al. 2011), and ReMap (Griffon et al. 2015). Gene expression data can be mined using resources such as Gene Expression Omnibus (Barrett et al. 2013), ImmGen (Heng, Painter, Immunological Genome Project Consortium 2008), and BioGPS (Wu et al. 2009, 2013). Likewise, databases of expression quantitative loci (eQTLs) (Carithers et al. 2015; Pickrell et al. 2010; Yang et al. 2010) can also be used to narrow down variants to those that affect the expression of a gene as a function of genotype. By incorporating cell types relevant to the disease or phenotype of interest, possible impacted regulatory mechanisms can be identified by intersecting the genomic coordinates of each variant with each of these datasets. Online tools such as Haploreg (Ward and Kellis 2012a) and RegulomeDB (Boyle et al. 2012) can also be used for this purpose. These tools take lists of variants as input, and produce lists of potentially impacted regulatory mechanisms for each variant. Additional resources can also be used to establish the specific molecules whose binding might be influenced by a given variant. For TFs and RBPs, CisBP (Weirauch et al. 2014) and CisBP-RNA (Ray et al. 2013) are currently the most comprehensive databases of TF and RBP binding site motifs, respectively. Both servers house web-based tools for ranking TFs and RBPs for each variant. For

miRs, web servers such as miRbase (Kozomara and Griffiths-Jones 2014) can be used. Collectively, all of these analyses aim to rank non-coding variants based on their likelihood of impacting regulatory mechanisms in relevant cell types, and ideally produce a finite number of experimentally testable regulatory mechanisms disrupted by specific variants. If, for example, binding of a specific TF is predicted to be impacted by a given variant, follow-up experiments such as Electrophoretic Mobility Shift Assays (EMSAs) (Staudt et al. 1986; Singh et al. 1986), ChIP-qPCR (need a citation), and reporter assays (Fort et al. 2011; Solberg and Krauss 2013; Rahman et al. 2001) can then be used for functional validation (e.g., (Harismendy et al. 2011; Kottyan et al. 2015; Musunuru et al. 2010; Lu et al. 2015; Pomerantz et al. 2009; Claussnitzer et al. 2015)).

14.8 Identification of Clinically Relevant Variants

In a clinical setting, the success rate of identifying a clinically relevant variant is between 20 and 30 %. The success rate depends on prior knowledge of the genetics of the disease and careful screening of the patient's clinical record and family history. Often, analysis will identify either variants of unknown significance (VUS) or a gene of unknown significance (GUS). Determining if these results are clinically reportable is difficult. Many institutions have developed expert panels of clinicians, geneticists and basic science researchers to carefully evaluate the potential of the variant to cause disease and determine if the evidence is strong enough to report the results to the patient. While VUS and GUS are often frustrating in a clinical setting, they are of great interest in a research setting. Often these results identify important genes for understanding the pathways that lead to disease, providing new research opportunities.

14.9 Conclusions and Future Directions

Next-generation sequencing has fundamentally altered the scope and nature of genetic questions that can be addressed in both the clinic and basic science research. Since the advent of NGS, advances have brought steadily increasing throughput and decreasing cost, to the point that these experiments are now affordable and easily accessible. However, the widely available technologies remain "short read" sequencers, with reads of a few hundred nucleotides at most. These reads can not be accurately aligned within large duplications or other highly repetitive regions; therefore, genetic variation within such regions can not be assessed with this technology. These and other limitations can only be overcome with long sequencing reads with low error rates. Functional interpretation of regions of genetic variation has proven to be an even larger challenge, highlighting the need for more research aimed at the functional characterization of genomic elements.

References

1000 Genomes Project, Abecasis GR, Adam A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Hyun Min K, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65. doi:10.1038/nature11632.

1000 Genomes Project, Adam A, Brooks LD, Durbin RM, Garrison EP, Hyun Min K, Korbe JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. doi:10.1038/nature15393.

Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011;8(1):61–5. doi:10.1038/nmeth.1527.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. doi:10.1016/S0022-2836(05)80360-2.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, et al. NCBI GEO: archive for functional genomics data sets – update. Nucleic Acids Res. 2013;41(Database issue):D991–95. doi:10.1093/nar/gks1193.

Bentley DR, Swerdlow HP, Shankar B, Smith GP, John M, Brown CG, Hall KP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53–9. doi:10.1038/nature07517.

Boyle AP, Hong EL, Manoj H, Yong C, Schaub MA, Maya K, Karczewski KJ, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22(9):1790–7. doi:10.1101/gr.137323.112.

Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. Biopreserv Biobank. 2015;13(5):311–9. doi:10.1089/bio.2015.0032.

Chadwick LH. The NIH roadmap epigenomics program data resource. Epigenomics. 2012;4(3):317–24. doi:10.2217/epi.12.18.

Chaisson MJP, Wilson RK, Eichler EE. Genetic variation and the De novo assembly of human genomes. Nat Rev Genet. 2015;16(11):627–40. doi:10.1038/nrg3933.

Claussnitzer M, Dankel SN, Kyoung-Han K, Gerald Q, Wouter M, Christine H, Viktoria G, et al. FTO obesity variant circuitry and adipocyte browning in humans. N Engl J Med. 2015;373(10):895–907. doi:10.1056/NEJMoa1502214.

Danecek P, Adam A, Abecasis GR, Albers CA, Eric B, DePristo MA, Handsaker RE, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8. doi:10.1093/bioinformatics/btr330.

DePristo MA, Eric B, Ryan P, Garimella KV, Maguire JR, Hartl CL, Philippakis AA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8. doi:10.1038/ng.806.

Eid J, Adrian F, Jeremy G, Khai L, John L, Geoff O, Paul P, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133–8. doi:10.1126/science.1162986.

Exome Aggregation Consortium, Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. BioRxiv. 2015;030338. doi:10.1101/030338.

Fort A, Fish RJ, Catia A, Roland D, Axel V, Marguerite N-A. A liver enhancer in the fibrinogen gene cluster. Blood. 2011;117(1):276–82. doi:10.1182/blood-2010-07-295410.

Griffon A, Quentin B, Jordi D, van Helden J, Salvatore S, Benoit B. Integrative analysis of public ChIP-Seq experiments reveals a complex multi-cell regulatory landscape. Nucleic Acids Res. 2015;43(4):e27. doi:10.1093/nar/gku1280.

Harismendy O, Dimple N, Xiaoyuan S, Rahim NG, Bogdan T, Nathaniel H, Bing R, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-Γ signalling response. Nature. 2011;470(7333):264–8. doi:10.1038/nature09753.

Heng TSP, Painter MW, Immunological Genome Project Consortium. The immunological genome project: networks of gene expression in immune cells. Nat Immunol. 2008;9(10):1091–4. doi:10.1038/ni1008-1091.

Hindorff LA, Praveen S, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106(23):9362–7. doi:10.1073/pnas.0903103106.

Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12:656–64.

Kottyan LC, Zoller EE, Jessica B, Xiaoming L, Kelly JA, Rupert AM, Lessard CJ, et al. The IRF5-TNPO3 association with systemic lupus erythematosus has two components that other autoimmune disorders variably share. Hum Mol Genet. 2015;24(2):582–96. doi:10.1093/hmg/ddu455.

Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42(Database issue):D68–73. doi:10.1093/nar/gkt1181.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980–5. doi: 10.1093/nar/gkt1113.

Langmead B, Cole T, Mihai P, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25.

Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013;152(6):1237–51. doi:10.1016/j.cell.2013.02.014.

Li H, Durbin RM. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25(14):1754–60. doi:10.1093/bioinformatics/btp324.

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. SNP detection for massively parallel whole-genome resequencing. Genome Res. 2009;19(6):1124–32. doi:10.1101/gr.088013.108.

Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, Shin H, et al. Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol. 2011;12(8):R83. doi:10.1186/gb-2011-12-8-r83.

Lu X, Zoller EE, Weirauch MT, Zhiguo W, Bahram N, Williams AH, Ziegler JT, et al. Lupus risk variant increases pSTAT1 binding and decreases ETS1 expression. Am J Hum Genet. 2015;96(5):731–9. doi:10.1016/j.ajhg.2015.03.002.

Maurano MT, Richard H, Eric R, Thurman RE, Eric H, Hao W, Reynolds AP, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190–5. doi:10.1126/science.1222794.

Musunuru K, Alanna S, Maria F-K, Lee NE, Tim A, Sachs KV, Xiaoyu L, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010;466(7307):714–9. doi:10.1038/nature09266.

Paul DS, Soranzo N, Beck S. Functional interpretation of Non-coding sequence variation: concepts and challenges. BioEssays: News Revs Mol Cell Dev Biol. 2014;36(2):191–9. doi:10.1002/bies.201300126.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Everlyne N, Jean-Baptiste V, Matthew S, Yoav G, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464(7289):768–72. doi:10.1038/nature08872.

Pomerantz MM, Nasim A, Li J, Paula H, Verzi MP, Harshavardhan D, Beckwith CA, et al. The 8q24 cancer risk variant Rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet. 2009;41(8):882–4. doi:10.1038/ng.403.

Rahman M, Hirabayashi Y, Ishii T, Kodera T, Watanabe M, Takasawa N, Sasaki T. A repressor element in the 5′-untranslated region of human Pax5 exon 1A. Gene. 2001;263(1–2):59–66.

Ray D, Hilal K, Cook KB, Weirauch MT, Najafabadi HS, Xiao L, Serge G, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013;499(7457):172–7. doi:10.1038/nature12311.

Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, et al. ENCODE data in the UCSC genome browser: year 5 update. Nucleic Acids Res. 2013;41(Database issue):D56–63. doi:10.1093/nar/gks1172.

Rothberg JM, Wolfgang H, Rearick TM, Jonathan S, William M, Mel D, Leamon JH, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475(7356):348–52. doi:10.1038/nature10242.

Singh H, Sen R, Baltimore D, Sharp PA. A nuclear factor that binds to a conserved sequence motif in transcriptional control elements of immunoglobulin genes. Nature. 1986;319(6049):154–8. doi:10.1038/319154a0.

Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7.

Solberg N, Krauss S. Luciferase assay to study the activity of a cloned promoter DNA fragment. Methods Mol Biol. 2013;65–78. (Clifton, N.J.) 977 (Chapter 6). Totowa: Humana Press. doi:10.1007/978-1-62703-284-1_6.

Staudt LM, Singh H, Sen R, Wirth T, Sharp PA, Baltimore D. A lymphoid-specific protein binding to the octamer motif of immunoglobulin genes. Nature. 1986;323(6089):640–3. doi:10.1038/323640a0.

Tennessen JA, Bigham AW, O'Connor TD, Wenqing F, Kenny EE, Simon G, Sean MG, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64–9. doi:10.1126/science.1219240.

Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012a;40(Database issue):D930–34. doi:10.1093/nar/gkr917.

Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. Nature Biotechnol. 2012b;30(11):1095–106. doi:10.1038/nbt.2422.

Weirauch MT, Ally Y, Mihai A, Cote AG, Alejandro M-M, Philipp D, Najafabadi HS, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158(6):1431–43. doi:10.1016/j.cell.2014.08.009.

Wu C, Camilo O, Jason B, Marc L, James G, Serge B, Hodge CL, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol. 2009;10(11):R130. doi:10.1186/gb-2009-10-11-r130.

Wu C, Macleod I, Su AI. BioGPS and MyGene.Info: organizing online, gene-centric information. Nucleic Acids Res. 2013;41(Database issue):D561–5. doi:10.1093/nar/gks1114.

Yang T-P, Claude B, Montgomery SB, Dimas AS, Maria G-A, Stranger BE, Panos D, Dermitzakis ET. Genevar: a database and java application for the analysis and visualization of SNP-gene associations in eQTL studies. Bioinformatics. 2010;26(19):2474–6. doi:10.1093/bioinformatics/btq452.

Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24(R1):R102–10. doi:10.1093/hmg/ddv259.

Chapter 15
Precision Pediatric Genomics: Opportunities and Challenges

Kristen L. Sund and Peter White

Abstract Precision medicine holds substantial promise for moving beyond the treatment of typical patients to the development and application of evidence-based precision care. As the nation embraces this new healthcare model, we emphasize the importance of utilizing phenotypic and genomic data of increasing volume and variety for accelerating disease discovery, translational science, and individualized care of patients. This requires a commitment that incorporates specific researcher, care provider, and disease-focused pursuits within a larger community of coordinated practice across and between each care institution. The change required to ensure a sound future for genome-centered care encompasses technical, data, behavioral, and organizational practices. Within pediatric institutions, enterprise-level commitment will be increasingly required to ensure that caregivers, researchers, patients and families, and patient support systems are sufficiently literate and invested in the promise of genomic medicine. We review here a number of successful and emerging trends in developing and implementing genomic practice at the level of an organization. These include broad patient ascertainment, consent, and regulatory practices; collection, representation, harmonization, and responsible sharing of genomic and observational data for large patient populations, including through electronic health record systems; rapid and robust analysis of genomic data for discovery and clinical decision-making; and empowering stakeholders to most effectively make use of newly generated genomic knowledge. We also discuss the importance of developing social structures that combine and maximize awareness, learning, and participation for genomic medicine. We include descriptions of the Center for Pediatric Genomics at Cincinnati Children's Hospital and the multi-institutional Genomic Research and Innovation Network as illustrations of enterprise-level genomic programs. Institutional commitments to integrated genomics practice will ensure the progression of genome-based pediatric practice, as well as deeper insights into the molecular basis of complex childhood disorders.

Keywords Genomics • Phenotype • Precision medicine • Genomics infrastructure • Data sharing

K.L. Sund (✉) • P. White
3333 Burnet Ave, Cincinnati, OH 45229, USA
e-mail: Kristen.Sund@cchmc.org; Peter.White@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_15

15.1 Existing Models for Broad Patient Enrollment

Much focus in the "big data" era involves technology and analytical methods. An additional and often overlooked challenge for establishing and generalizing genomic and precision medicine studies is the development of robust and scalable regulatory and compliance processes. Precision medicine initiatives across the country use a variety of models for consent and enrollment of participants. Such efforts require balancing the desire to facilitate the efficiency and quality of research with the proper protection of subjects. Solutions are often complicated by a wide range of local, state or district, federal, and international regulatory practices that are frequently evolving, ambiguous, and occasionally in conflict. Current practices range from a resource-intensive full, in-person consent, to an intermediate mail-in consent, to a revised "consent to treat." Ethical enrollment needs to incorporate assurance that participants sufficiently understand a study, which often incurs a substantial educational effort for genomic studies. To assist with such costs, innovative methods of implementation are developing around electronic consent processes and patient-centered portals. Multimedia tools and technologies include interactive computer and touch-screen presentations, take-home audiotape supplements, video vignettes, and powerpoint slideshows (Nishimura et al. 2013).

Two major institutional projects illustrate successful paths to broad patient enrollment, phenotype documentation, and sample acquisition. Geisinger Health System has a partnership with Regeneron Pharmaceuticals to generate sequencing data for up to 250,000 individuals combined with electronic medical record data for participants, as part of Geisinger's MyCode Community Health Initiative. (Carey et al. 2016). MyCode enables study participants to enroll through a prospective, in-person consent process. Participants contribute biomaterials for a centralized biobank and provide permission to link these samples to phenotypic data collected in the Hospital's centralized electronic health record (EHR). MyCode samples can be linked to clinical EHR data through a data broker who can access identified data. As of September, 2015, more than 90,000 patients had enrolled in the biobank, including 3600 pediatric patients that were enrolled through parental consent. Patients contribute a median of 60 clinical visits and 12 years of phenotypic data from the EHR. This data resource often requires extensive mining for re-use, including the use of natural language processing. For this work, Geisinger has leveraged efforts for developing methods for computable phenotyping, as supported by the NIH's Electronic Medical Records and Genomics (eMERGE) Network (Gottesman et al. 2013). To extract a cohort, custom-validated phenotypic algorithms are designed for clinical traits of interest and extracted using the Phenotype KnowleledgeBase (PheKB) (Rasmussen et al. 2014). Inclusion criteria are defined by extracting phenotypes from billing codes (CPT, ICD-9), laboratory results, and vital signs. Participants enrolled in MyCode agree to be re-contacted for clinically actionable results, and examples of the types of results that will or will not be returned are provided to subjects and/or their parents. An oversight committee has been established to determine which clinically validated results should be returned and the

most appropriate process for returning the information. Clinically actionable results are added to the patient electronic chart, and the patient is provided subsequent access to genetic counselors and other experts to discuss findings.

The Mayo Clinic is a nonprofit healthcare organization engaged in medical care, research and education throughout the US. Mayo's Center for Individualized Medicine aims to bring the latest discoveries from research to the clinic in the form of new genomics-based tests and treatments. Mayo has established a large collection of biological samples with associated patient-reported health information and electronic medical record health information (Ridgeway et al. 2013) . This was accomplished by recruiting patients who responded to a mailed consent. Patients scheduled for primary care appointments are randomly recruited by mail and through recruitment desks in two locations. For mail enrollment, completed materials can be returned via an enclosed postage paid envelope. The Mayo clinic has started to enroll healthy patients in a whole genome sequencing protocol in partnership with Helix (more on page 305).

Among Children's Hospitals, two are particularly notable for having invested in institutional approaches to broad patient enrollment. The Center for Applied Genomics at the Children's Hospital of Philadelphia (CHOP) is a long-established program that pioneered the idea of broad consent for genomic studies. CHOP has established a patient recruitment network across its busiest healthcare intake facilities that utilizes research coordinators for consenting both healthy and unwell children and parents. Subjects are asked for a short medical history synopsis, and their EHR data is linked to blood and DNA samples in a central warehouse in a de-identified manner. This resource, which encompasses nearly 400,000 samples, has been used for genome association (GWAS) and sequencing studies for many disease projects, including obesity, diabetes, autism, ADHD, and asthma, and as a healthy control cohort for analysis of structural variation (Bonnelykke et al. 2014; Bradfield et al. 2012; Elia et al. 2010; Gai et al. 2012; Hakonarson et al. 2007; Shaikh et al. 2009). Finally, in support of their institutional biobank of 800,000 samples and their enterprise genomic efforts (see below), Cincinnati Children's Hospital has instituted an opt-in consent for broad use of residual blood samples that is incorporated into the patient registration process. Consents and samples are linked to the de-identified institutional data warehouse, which is accessible for research studies with IRB authorization and an honest broker policy. These efforts have led to a number of scientific advances, notably including GWAS-based discoveries for a number of diseases through the NIH's eMERGE (Bush et al. 2016; Gottesman et al. 2013; Hall et al. 2015; Namjou et al. 2013, 2015).

15.2 Phenotypic Data Standardization

A paradox of precision medicine is that although the focus is on the individual patient, deriving confidence in an individualized treatment plan requires sizable patient populations, in order to obtain statistical rigor for stratifying subpopulations

in terms of onset, course, outcome, or intervention. Moreover, requirements for patient and data volumes increase dramatically for complex genetic disorders. As genomic sequencing costs continue a trend towards affordability, a number of projects have been initiated to generate genomic datasets for large cohorts. These include the ExAC consortium (90,000 existing whole exome sequences), and plans for the National Heart, Lung, and Blood Institute's Genome Sequencing Program (70,000 subjects), the UK Biobank (500,000 subjects), the Million Veteran Program (1 M), the NIH's Precision Medicine Initiative (PMI) (1 M), and a similar PMI initiative soon to be launched in China. While tremendous in scale and opportunity, this nature of genomic effort generally does not collect phenotypic data of suitable granularity for many pediatric disorders, due to disorder rarity and variability in presentation. Therefore, for genetic diseases of childhood, precision medicine often requires both large-scale and highly precise collection of phenotypic data. Especially for rare disorders, this collection process typically requires the formation of pediatric-focused data networks (see Chap. 10), where multiple institutions together share this data and collaborate on a collective study or output. Data sharing across institutions introduces the need for data representation and data quality standards, in order to ensure interoperability, reproducibility, and effective re-use. The granularity, completeness, expansiveness, and level of rigor needed to properly conduct these studies often exceeds what is available in EHR data, which can introduce a need for ancillary collection of additional data to support specific avenues of discovery. However, patient clinics are busy and overloaded clinicians are not traditionally focused on or incentivized for augmenting phenotypic data for research evidence-based medicine studies. Much of the data in the clinical chart that is useful for research studies is typically in unstructured clinician notes. Further, customized data capture systems are expensive, poorly generalizable, and often challenging to implement across multiple institutions. The ideal is to improve the quality of data input in the clinic, but this will first require optimization of EHR phenotyping tools and large-scale culture change (learn more in Chap. 10).

As global solutions for standardized data capture and representation continue to evolve, local data representation efforts relevant to genome-based phenotyping are emerging. One important new standard is The Human Phenotype Ontology (HPO) (Kohler et al. 2014). This project was established to create a comprehensive structured nomenclature around abnormal human phenotypes, to facilitate communication about the phenotypes and associated genetic findings, and to establish consistency in disease representation. The HPO initiative has built upon many years of ontology efforts used for phenotype-based characterization of several model organisms, which was pioneered and led by the Gene Ontology Consortium (Gene Ontology 2015). While initial concentration was on dysmorphologies, HPO has more recently grown to include a more complete representation of organ systems, tissues, and disease states. In 2014, HPO's data dictionary included 10,088 terms that describe human abnormalities and disease. About half of these terms are also associated with more extensive descriptions on affected cell type, function, embryology, and pathology. These descriptions can be used with phenotypic data in other model organisms through Gene Ontology and organism-specific term mappings.

HPO is structured as a directed acyclic graph, allowing for parent and child term groupings. Terms are defined and tied to known associated genes and diseases through annotations that include Online Mendelian Inheritance in Man (OMIM) and DECIPER (Amberger et al. 2015; Bragin et al. 2014), MySQL, and/or other web-based tools. A number of disease or function-focused ontologies have adopted HPO as a framework to promote semantic interoperability. As an example, The Monarch Initiative (Mungall et al. 2015) contributes another extension to HPO by creating semantic similarity algorithms across species for assessment of genetic variants and phenotypes. The Monarch web portal allows for upload and filtering of whole exome data with immediate comparison to known animal phenotypes and genotypes for diagnosis and disease mechanism research. Multiple groups are working to optimize the availability and use of HPO language through automation, adoption into clinical flows, and incorporation into commercial tools (Hamosh et al. 2013; Girdea et al. 2013).

15.3 Genomic Data Standardization

In order to realize the potential of genomic variation underlying human disease, large datasets will have to be processed and analyzed in unison, particularly for interpretation of rare variants. As technology has advanced and data processing costs continue to exponentially decrease, the major challenge has shifted from generation to interpretation of genomic data. To begin the process, standardized phenotypes are compared to variants identified through targeted, whole exome, or whole genome sequencing to look for associations between phenotypes and genotypes. However, wet and dry lab practices vary across the world, and the large scale data from different sources will not be comparable without measures to align data processing and annotation for optimized interrogation.

The Global Alliance for Genomics and Health (GA4GH) (Lawler et al. 2015) was created to develop standards for broad data sharing around the world. This team recognizes that the future of genomic interrogation lies in the interpretation of large data sets—and that this effort will require interoperability between dataset nomenclature and annotation. The Global Alliance Data Working Group created a free Application Programming Interface (API) to facilitate exchange of next generation sequencing data across diverse organizations and platforms. In addition to data standardization, the API provides variant annotation and genotype/phenotype association data. GA4GH has also participated in work with the Matchmaker Exchange (Buske et al. 2015; Philippakis et al. 2015) to create a federated platform for matching genomic variants and phenotypes with disease gene experts across disparate sites, in order to establish a distributed interpretation service. A number of clinical genetics laboratories participate in Matchmaker Exchange around the world. This network provides improved annotations for known and uncertain variants. These annotations can then be added to public annotation datasets such as ClinGen, ClinVar, dbSNP, and COSMIC.

The GA4GH recognizes that the challenges associated with genomic and health information data sharing are not purely technical. To address the reluctance of some investigators to share data, one of the first initiatives of GA4GH was to create a "Framework for responsible sharing of genomic and health-related data (Knoppers 2014)." The purpose of this document is to foster responsible data sharing through (1) respect for individuals, families and communities; (2) the advancement of research and scientific knowledge; (3) the promotion of health, well-being and equality; and (4) fostering trust, integrity and reciprocity. The core elements of the framework are transparency, accountability, engagement, data quality/security, privacy/data protection/confidentiality, risk-benefit analysis, recognition/attribution, sustainability, education and training, accessibility, and dissemination. The framework is designed for community-wide adoption, irrespective of GA4GH membership.

Data standardization is also a key focus for a public-private-academic collaboration known as the The Genome in a Bottle Consortium (Zook et al. 2014). The partnership is hosted by the National Institution of Standards and Technology (NIST). The primary goal of this Consortium is to develop technical infrastructure for implementation of genome sequencing into clinical practice. To do so, the Consortium is putting into practice a set of genomic technical and methodology reference standards, and making reference genomic data available to the public. These standards have been adopted by many clinical and research laboratories, resulting in a substantial increase both in site-specific quality, and also in participatory improvement of best practices. The Genome in a Bottle Consortium has also formed an analysis group to develop high quality phased variant calls for others to use as a benchmark for their data processing pipelines. The references will be able to identify potential biases in sequencing methods and give users the ability estimate the confidence of reported variant calls. The group's focus includes overcoming specific challenges in the areas of short read assembly, structural variants and phasing. As each of these challenges are addressed, the updated infrastructure is made available to the public.

15.4 Federated Genotype/Phenotype Searches for Cohort Expansion

Data fragmentation has made the task of cohort expansion challenging, particularly in rare disease. A major challenge is the lack of open data exchange, as such data is usually sequestered in individual laboratories, focused disease networks, and institutions. This is confounded by issues of patient consent and data ownership. The Matchmaker Exchange (MME) (Philippakis et al. 2015) was created to address these challenges by establishing a network of genomic/phenotypic data sets that are accessible through a central API. To optimize institutional regulatory infrastructure and maintain the ability to obtain longitudinal data, the collaborative group opted to

build a federated network that is connected through a common API. Programs that wish to participate must:

- Require users to submit case level data
- Establish at least two point-to-point API connections
- Contain useful genomic/phenotypic data content
- Implement algorithms for variant matching
- Enable dual notification of data requestor and data depositor, including sharing user identities and contact information
- Submit disclaimers for the open-source code repository to GitHub
- Store queries over time for auditing purposes
- Meet data and infrastructure security criteria

The API facilitates queries that are executed in the data stores of participating institutions, along with the response that provides information about potential matches (Buske et al. 2015). To standardize terminology, the MME API uses HPO terminology (Kohler et al. 2014) for phenotype representation and Sequence Ontology (Eilbeck et al. 2005) terms for genotype. Current efforts revolve around simple matching for cohort expansion of rare disease based on a common phenotype or genotype. The long-term goal is to transition to a model that allows one-sided hypothesis matching that enables a single submitted variant to be compared to previously submitted and broader datasets, such as total variant sets (VCFs) from one or more additional individuals of interest. The MME is an open-source example of the changing environment in big data science that will make possible broad-scale variant data sharing.

15.5 Genomic Data Sharing

Historically, the size and complexity of pediatric datasets has been substantially limited by data attainment costs and single investigator-focused academic models. These constraints have been largely eliminated through technology advances, as well as the realization that the genetics of most disorders is multifactorial. As molecular data acquisition costs continue to fall, the field is quickly accumulating large data sets that overcome localized informatics and IT capacities, both in terms of IT infrastructures and sufficient knowledgeable personnel. Large-scale genomics initiatives are thus following the lead of other data-intensive scientific disciplines such as meteorology, climate science, and astrophysics. Through the concepts of team science and distributed infrastructures, these initiatives have transitioned from local resources to shared data and elastic compute facilities, often referred to as the "cloud". Well-executed cloud storage and computing provides several opportunities for accelerating high-quality science, but it also challenges existing researcher and institutional behavior paradigms. A cloud architecture provides the means for readily establishing a disease or project-focused commons for collaborating institutions to store data, develop and execute applications, and share both data and results. This

arrangement benefits scientific discovery by sharing and economizing infrastructure costs, as well as by building trust through facilitated data sharing and dissemination. The growing advent of open data initiatives is illustrative of the attraction to such models. Cloud infrastructures address many performance questions by offering seemingly unlimited, on-demand storage and compute resources that are connected through high-speed networks. Thus, the cloud offers a model for scalability both in terms of data storage and compute capacity, while allowing participating institutions autonomy to make their own analysis workflow decisions. However, the cloud environment also represents new challenges to genomic data sharing from legal, regulatory, behavioral, and security perspectives (Dove et al. 2015). Potential hidden costs of cloud architectures include perceptions of inappropriate data use, the need for compromise in choosing appropriate technology and analytical strategies, and short-term costs required to convert local workflows to a distributed computing environment. Additionally, conversion to a cloud environment can create service provider dependencies, as well as increasing the need for well-defined data, intellectual property, and academic credit governance. Nevertheless, a number of genomic initiatives are moving to a cloud strategy. These early adopters are typically consortium-based projects that can leverage a strong central incentive for participation. In this section, we focus on one phenotype—autism spectrum disorder—and the work being done in this disease to make genomic data more accessible to researchers.

Autism spectrum disorder (ASD) is a collective neurodevelopmental diagnosis that includes manifestations of impaired social interaction and communication skills. The diagnosis is made in 1:500 children— these children meet criteria set forth in the diagnostic and statistical manual of mental disorders, 5th edition (DSM-5). The etiology of ASD is not well understood, but it is likely due to some combination of environmental and genetic factors, and there is substantial evidence supporting a sizable heritable component (Hallmayer et al. 2011). Organizations that support ASD research have invested significantly in whole genome surveillance, which has led to indications of genomic complexity (Gai et al. 2012; Geschwind and State 2015). In particular, parallel initiatives organized by the advocacy groups Autism Speaks and the Simons Foundation are coalescing their autism genomic research infrastructures using a cloud strategy. Both organizations have recently partnered with cloud providers to share large-scale genomic data with researchers.

MSSNG is a partnership between Autism Speaks and Google Genomics to sequence 10,000 individuals from the NIH-sponsored Autism Genetic Research Exchange Repository, itself a consortium designed to share genomic data from families with a member affected with ASD. The effort aims to provide sequences, annotations, and phenotype data to autism researchers through the Google Cloud Platform (Yuen et al. 2015). Genomic efforts of this magnitude, which generate petabyte scale data and enormous computational requirements for iterative analysis, outstrip the ability of any single pediatric institution to support. Google Genomics created an API under the standards of GA4GH, which can be used to access data for the MSSNG project. At the same time, MSSNG has created a data use agreement that creates an environment for responsible data sharing. Participating users must

agree to a number of shared governance provisions, including use restricted to research; and compliance with MSSNG's privacy, intellectual property, academic credit, and compliance policies.

The Simons Foundation Autism Research Initiative began to explore the genetics of autism spectrum disorders through the collection of copy number variation data generated by microarray-mediated genome surveillance (Fischbach and Lord 2010). The organization subsequently supported the generation of genomic sequencing data over time. The current Simon's data resource contains sequence and associated phenotypic data from 2600 quadplex families with a child who has a diagnosis of autism spectrum disorder, as well as an unaffected sibling to serve as an internal control. The data can now be accessed through a commercially-supported platform after researchers sign a data agreement similar in scope to the MSSNG governance policy. The Simons data resource includes applications for data processing, including GATK (Van der Auwera et al. 2013) and FreeBayes.

15.6 Implementing Genomics in the Clinic

New focus and investment in precision medicine is re-emphasizing the need to make genomic discoveries translatable. The implementation of genomic information into clinical work flows will require a transformation in culture and training, data infrastructure, and patient education. A number of existing efforts are exploring the burden and developing methods to overcome these barriers. The NIH-supported MedSeq (Vassy et al. 2014) and BabySeq (Frankel et al. 2016) projects are randomized, controlled trials that simultaneously enroll patients and their physicians in a comparison of the current standard of care to the standard plus the addition of whole genome sequencing results. Babyseq, modeled after MedSeq, enrolls newborns and their parents in both standard, healthy settings at an adult hospital or in the context of a neonatal intensive care unit at a children's hospital. Subjects are randomly assigned to receive standard newborn screening results either with or without genome sequencing results. Parents return after 6 weeks for a disclosure interview where they receive results related to childhood-onset conditions, carrier status, and pharmacogenomics. Participants are asked to complete surveys about their experience at enrollment, disclosure, and again 3 and 10 months after disclosure. Both the BabySeq and MedSeq projects are building communities of genomics practice by educating patients/families and clinicians in a variety of settings. These projects also build trust with their patients and the broader community, a necessary step in the implementation of new technologies.

Another NIH-sponsored initiative, Implementing GeNomics In PracTicE (IGNITE) (Weitzel et al. 2016), is considering a variety of data types in planning for genomic implementation. IGNITE comprises six projects that are creating genomic practice models for results dissemination in electronic health records, with EHR-enabled clinical decision support (CDS). Current projects aim to explore genetic markers for disease risk prediction and prevention, develop tools for family history data in diverse settings, incorporate actionable pharmacogenomics data into clinical

care, refine diagnostics through mutational analysis, and develop new educational approaches. The network collects data on clinical indications, family history, genetics and outcomes, including individual patient genetic and pharmacogenetics test results. Genomic data collected in the IGNITE studies is expected to inform patient care decisions, such that all genetic testing is carried out in a CLIA certified clinical laboratory. The IGNITE goal is to link outcomes to genomics-informed clinical decision-making to determine how this information improves patient care. While nascent, this strategy represents a promising future direction for personalized medicine.

The US Government's Health Information Technology for Economic and Clinical Health (HITECH) Act included a Meaningful Use incentive that led to the broad adoption of EHR systems. While the use of electronic systems is leading to the ability to transmit health information across institutions, substantial challenges in data formatting and implementation remain, and the readiness of healthcare organizations to return genomic results is limited (Tarczy-Hornoch et al. 2013). If executed properly, genetic and genomic results collected in the EHR will be available for use in precision medicine and research, though each area faces different challenges (Marsolo and Spooner 2013). Two major NIH-funded initiatives, eMERGE (introduced on p. 296) and the Clinical Sequencing Exploratory Research (CSER) program have recently collaborated to assess the current state of inclusion of genetic information in the EHR and to envision a potential future state (Shirts et al. 2015). Not surprisingly, these groups indicated substantial heterogeneity in the methods used to integrate genetic data into the EHR, and the location and accessibility of the data in the record. For the majority of genetic data types, more than half of the reporting institutions replied that the genetic data was stored as text blobs, such as in PDF format. Many institutions report the presence of genetic information in multiple places in the EHR record, and the majority reported that the source laboratory was the largest determining factor of information location. Interestingly, only 42 % of institutions reported a centralized effort to consolidate genetic information in the EHR. The CSER-eMERGE EHR Integration Working Group proceeded to prioritize a list of 20 recommendations for improvement of methods to store genomic data in the EHR, and those ranked highest were all related to improving clinical decision support. As most data warehouses pull from discrete EHR data, it will be crucial to find ways to organize genomic results data and make it queryable both for research and precision medicine care. The level of access required is also under discussion, as current EHR systems are not equipped to handle large processed or raw genomic data.

15.7 Towards Applications for Consumer Genomics

As technical capabilities evolve to more readily accommodate genomic data, many companies are exploring models of consumer-driven, consumer-oriented use of genome results. These companies are implementing patient-facing applications to

return genomic results, share genomic data, and engage in research opportunities. It is now possible for a consumer to have their whole genome sequenced and to have results returned to their mobile device without a clinical entity serving as an intermediary (direct-to-consumer (DTC) genetic testing). DTC genetic companies are likely to accelerate the exposure of the public to genomic testing through ease-of-use tools. Interestingly, customers will form opinions—good or bad—about genomic medicine based on their experiences with DTC vendors. There is great potential to gain or lose community trust around these technologies if they are not sufficiently validated, do not provide reliable, useful information to the client, or do not provide good customer support when unexpected genetic results are returned. Recently in the US, in response to concerns regarding the value and accuracy of DTC genetic testing, the Food and Drug Administration has begun to regulate private service providers. The proper balance between consumer choice and the need for clinical oversight of genomic results interpretation is a topic of considerable current debate (Evans et al. 2010; Green and Farahany 2014)

An increasing number of established companies and new start-ups are entering the market to utilize personalized devices for genomics test results, or for participating in research. In the beginning of 2016, the company 23andMe partnered with Apple to develop a consumer data application to share genomic and health information data with researchers through an iPhone application (23andme 2016). This partnership includes use of Apple's ResearchKit development platform that is designed to accelerate biomedical research. Veritas Genomics (Genomeweb 2016) is the first genome sequencing company that is offering direct-to-consumer whole-genome sequencing and interpretation. Veritas delivers results of medically actionable genes though a digital report that can be explored with a dedicated application (Ray 2016). In this space, the challenges appear to be focused more around consumer understanding and consent than technical needs. Furthermore, the follow up care needed by these patients will likely occur in genetics specialty clinics that already have long waiting lists for patients. Additionally, many clinicians in these clinics are focused on Mendelian disease and may not have the answers desired by consumer driven genomics.

One potential pathway for empowering consumers while simultaneously providing appropriate clinical guidance is partnership between DTC providers and academic medical centers. To accelerate precision medicine through consumer adoption of sequencing, the Mayo Clinic recently partnered with Helix (Healthcare IT News 2015), a company affiliated with sequencing provider Illumina. Companies such as Helix are exploring ways for consumers to manage their own genomic data and utilize genomic data dissemination and interpretation applications, including mobile applications that are of interest to them. In the Helix model, industry and academic partners develop applications that provide data as well as responsibly educating, interpreting, and providing decision support regarding personalized genomic data, such as inherited risks for pediatric and adult disease, ancestry, fitness, and wellness. The Mayo Clinic/Helix partnership is a good attempt to cooperate to address many of the concerns introduced by consumer genomics and will likely yields guiding strategies for other centers.

15.7.1 Genomics in Pediatrics

Genomics—and subsequent integrative –omics— represents a paradigm shift in biomedical health practice as we know it. Precision genomics offers great promise but also requires a transformational shift in approach from the typical practice of treating a fictional "average" patient, to evidence-based medicine that precisely considers factors derived from each individual. In order to rapidly and broadly implement precision genomics, substantial changes from current models in institutional culture and the practice of medicine are necessary. As precision genomics touches all broad areas of biomedical research and practice, it will be increasingly important to develop informed participants and practices whose knowledge and skills are well aligned and coordinated. From a pediatrics perspective, this is confounded by a challenging ethical and regulatory framework, as well as by the typically strong roles of parental and community stakeholders. Moreover, decisions made by parents today about genetic information may impact children—who may or may not be involved in shared decision-making—for a lifetime. In the research realm, academic models often incentivize single investigator discovery and translation at the expense of team science, which is critical for interpreting genomic data and making significant advances towards understanding multifactorial diseases. Cincinnati Children's is an example of an academic medical center that is carefully considering this landscape as they evolve their genomic strategy. Part of this strategy includes the creation of the Center for Pediatric Genomics (CpG) and leadership in the national pediatric genomics network, the Genomic Research and Innovation Network (GRIN). These initiatives will be used as general exemplars to illustrate major issues to be considered; other institutions have approached certain of these issues in slightly different ways.

The **Center for Pediatric Genomics (CpG)** was conceived as a construct within Cincinnati Children's combined clinical and research genomics community to accelerate the translation of genomic information into clinical care. CpG's vision is to create an institution-wide "community of practice" around genomics in order to effectively implement precision genomic medicine for improving the health of children. CpG's mission is to accelerate discovery and translate genomic knowledge into improved health at the enterprise level. Cincinnati Children's effort is novel as it uses a grassroots approach to encourage internal and external stakeholder engagement. CpG funds innovation; incubates and aligns resources; fosters collaboration; and provides a facilitating infrastructure for data, technology, analysis, and regulatory practices.

Community engagement is challenged by diversity in education, socioeconomic status, and health interests. To fully engage, institutions must be prepared to educate their patients, clinicians and researchers in a way that directly involves them in healthcare decisions regarding genomics. The CpG program focuses on reducing uncertainty around genomics from each type of stakeholder: for example, from patients and families about genomic results being communicated by a clinical lab; from participants asked to consent to a broad-based genomic study; from clinicians regarding how to interpret a genomic result, or when to order a particular test; or

from genomic researchers regarding how their expertise can best align with opportunities for improved health. One objective of this approach is to develop *trust* with families so they can make informed genome-based decisions about improving their health, as well as understanding the implication of their choices. CpG also provides patients and their families the *opportunity* to participate in research studies to improve their own care and the care of others. Studies and experience from CHOP, Cincinnati, Vanderbilt, and other pediatric institutions show that the majority of parents and/or children *will* participate when given the option (Desai et al. 2016; McGregor et al. 2013). To fully engage patients and families, the CpG program is transparent when it engages in research and education, including assurance that participants understand how their data will be used with and without their consent. After consent for broad data sharing activities, conversation acknowledges the inability to precisely describe future studies, but an overall description of the expected types of studies and the types of data sharing is shared. The concept of how sample use is optimized for best potential, and that study leaders are providing data in good faith to maintain participant trust for future studies is also communicated. Wherever possible, the CpG program continually engages participants in a cycle of genomic education, knowledge and engagement so that they can learn more as they understand more.

On another level, CpG attempts to engage clinicians and basic researchers in creative ways that are intended to induce positive culture change. The program places emphasis on facilitating the creation of collaborative teams of clinicians and bench researchers that invokes active communication, in order to understand the variety of languages, perspectives and contributions. The objective is to raise awareness of capability across dynamic teams in order to broaden the likelihood of innovative spark. As with the patient engagement initiative, this focus on developing a collaborative environment is intended to increase trust across the institution. One mechanism that the CpG program has found to foster collaboration is by funding innovative genomic projects that pair multidisciplinary teams of basic and translational scientists with practicing clinicians. This peer-reviewed pilot program has executed three rounds of pilot funding that has generated over 120 collaborative applications across nearly all clinical and basic science units at Cincinnati Children's. Funded principal investigators participate in a monthly seminar that brings together basic researchers, who may be struggling with their first Human Subjects protocol, with clinicians who might be learning the technicalities of applying next generation sequencing to a rare disease cohort. The CpG leadership also takes an active role in incubating both funded proposals and promising applications by providing consultation services, resources, matchmaking, and infrastructure for improvement. Another important aspect of the program is the development and coordination of educational programs designed both for clinicians and researchers, along with continual engagement of stakeholders through smaller personalized conversations where possible. CpG has recognized that clinician engagement is critical, as clinicians are typically the nexus for decisions regarding diagnostic and therapeutic approach, and also because these practioners can best provide the phenotypic data needed for precise genome interpretation.

In the first round of CpG pilot funding, genomic data sharing was quickly recognized as a considerable challenge within Cincinnati Children's. This led to investment in IT and informatics infrastructure for uniform genomic data acquisition, processing, and dissemination. The central facility for these activities is an application suite that provides sequence processing and annotation, sequence and genomic variant data management, and web-enabled query and reporting functions. This resource (¡VIVA!) serves as a community hub for institution-wide integration, processing, storage, and sharing of genomic data. The data resource uploads genomic data files from multiple sources, including individual investigators, research sequencing partners, clinically generated raw data files from consented participants, and publically available datasets. Once uploaded, data is processed using a standardized and documented suite of tools and workflows to ensure that the repository contains the most consistent, comparable data possible. Furthermore, ¡VIVA! serves as an institutional data asset that allows queries of aggregate genomic data, and under the proper consent, also allows patient level genomic data to be linked with phenotypic data. The tool provisions data access by principal investigator and his/her collaborators; with the proper consent, data can be shared across the institution. Importantly, data contributors are incentivized for data sharing through a mix of functional incentives. Finally, ¡VIVA! enables researchers to identify existing biobank specimens with a particular phenotype or genotype of interest. Taken together, this data resource provides Cincinnati Children's efficiencies both for catalyzing collaborative discovery and analysis within CpG, and for participation in external partnerships and grant opportunities. Challenges faced by the development of this infrastructure are similar to those faced in national and international genomics projects, including scope definition, properly incentivizing data sharing, ensuring proper balance between appropriate data privacy and broad use, and incorporating accurate and sufficiently granular phenotypic data.

The **Genomic Research and Innovation Network** (GRIN) is a new collaboration between three leading pediatric institutions: Boston Children's Hospital, Cincinnati Children's, and the Children's Hospital of Philadelphia. GRIN's vision is to accelerate genomic discoveries in pediatric populations through a collective ability to pursue well-designed, carefully selected research studies of genetic and genomic data associated with strongly characterized phenotypic information. This work aligns well with learning health systems such as PedsNet (Chap. 10). Combining genetic and health information from these pediatric institutions into a scalable, cloud-based, stakeholder-accessible infrastructure is expected to create an unparalleled capability in pediatrics. As partnerships with industry often accelerate the ability of participant data to have an impact on patient care, the GRIN leadership is considering a long term strategy to make the data resource available to both academic and industry partners through a co-governed entity known as the Pediatric Data Trust (PDT) under the appropriate participant consent.

The PDT is being architected as a *platform* that will facilitate the sharing of clinical and -omic data between participating institutions. In addition, the PDT provides facilitated *services* that can, with the appropriate approvals, support a variety of activities. These include data discovery, cohort identification and expansion (espe-

cially for rare disorders), and distributed data analyses. In addition, the PDT is expected to accelerate strategic observational, comparative effectiveness, and translational research projects. The PDT infrastructure is designed with three tiers of access. A public tier contains aggregate population-level statistics that allows potential users to understand the types of information included in the data resource. A privileged tier is a shared computation and data space where users with appropriate authorization can execute queries or perform collaborative analyses. Separate private tiers are controlled by each participating institution. Each private tier instance contains the entire staged data asset to be considered for possible inclusion in studies from the respective institution. The PDT also includes facilitators whose role includes the vetting and execution of data requests, where each GRIN member (and individual data contributor) can determine whether their data should be shared for a particular request.

15.8 Conclusion

The ideal of precision medicine in pediatrics is to provide the best possible care for a child based on our latest understanding of the science. A precision future means better treatment for sick kids. It means better understanding of the underlying causes of disease, with an emphasis on prevention where possible. Many of the initiatives outlined in this chapter demonstrate developing strategies to collect better, larger datasets to amass the data we need to generate answers about how to treat the individual child in the clinic at any given moment. Challenges remain in the collection of complete phenotypic data in a common language during routine clinical visits, the organization of genomic data in the electronic health record, the education of patients and clinicians alike, the culture of practice that leads to a cycle of educate, discover, translate and treat; but perhaps the greatest challenge is in understanding and interpreting these large data sets. The future will require novel analytic capabilities to transform massive datasets to meaning in the clinic. The incorporation of molecular data into standard care has a long way to go—but we are rapidly overcoming barriers to provide precision care for kids.

References

23andme. 23andme Enables Genetic Research for Researchkits Apps [Press release]. 2016. Retrieved from http://mediacenter.23andme.com/blog/researchkit/.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(Database issue):D789–98. doi:10.1093/nar/gku1205.

Bonnelykke K, Sleiman P, Nielsen K, Kreiner-Moller E, Mercader JM, Belgrave D, … Bisgaard H. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. Nat Genet. 2014;46(1):51–5. doi:10.1038/ng.2830.

Bradfield JP, Taal HR, Timpson NJ, Scherag A, Lecoeur C, Warrington NM, … Early Growth Genetics C. A genome-wide association meta-analysis identifies new childhood obesity loci. Nat Genet. 2012;44(5):526–31. doi:10.1038/ng.2247.

Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, Swaminathan GJ. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res. 2014;42(Database issue):D993–1000. doi:10.1093/nar/gkt937.

Bush WS, Crosslin DR, Obeng AO, Wallace J, Almoguera B, Basford MA, … Ritchie MD. Genetic variation among 82 pharmacogenes: the PGRN-Seq data from the eMERGE Network. Clin Pharmacol Ther. 2016;doi:10.1002/cpt.350.

Buske OJ, Schiettecatte F, Hutton B, Dumitriu S, Misyura A, Huang L, … Brudno M. The matchmaker exchange API: automating patient matching through the exchange of structured phenotypic and genotypic profiles. Hum Mutat. 2015;36(10):922–27. doi:10.1002/humu.22850.

Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, … Ledbetter DH. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet Med. 2016. doi:10.1038/gim.2015.187.

Desai A, Connolly JJ, March M, Hou C, Chiavacci R, Kim C, … Hakonarson H. Systematic data-querying of large pediatric biorepository identifies novel Ehlers-Danlos Syndrome variant. BMC Musculoskelet Disord. 2016;17(1):80. doi:10.1186/s12891-016-0936-8.

Dove ES, Joly Y, Tasse AM, Public Population Project in, G., Society International Steering C, International Cancer Genome Consortium E, … Knoppers BM. Genomic cloud computing: legal and ethical points to consider. Eur J Hum Genet. 2015;23(10):1271–78. doi:10.1038/ejhg.2014.196.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. The sequence ontology: a tool for the unification of genome annotations. Genome Biol. 2005;6(5):R44. doi:10.1186/gb-2005-6-5-r44.

Elia J, Gai X, Xie HM, Perin JC, Geiger E, Glessner JT, … White PS. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. Mol Psychiatry. 2010;15(6):637–46. doi:10.1038/mp.2009.57.

Evans JP, Dale DC, Fomous C. Preparing for a consumer-driven genomic age. N Engl J Med. 2010;363(12):1099–103. doi:10.1056/NEJMp1006202.

Fischbach GD, Lord C. The Simons simplex collection: a resource for identification of autism genetic risk factors. Neuron. 2010;68(2):192–5. doi:10.1016/j.neuron.2010.10.006.

Frankel LA, Pereira S, McGuire AL. Potential psychosocial risks of sequencing newborns. Pediatrics. 2016;137 Suppl 1:S24–9. doi:10.1542/peds.2015-3731F.

Gai X, Xie HM, Perin JC, Takahashi N, Murphy K, Wenocur AS, … White PS. Rare structural variation of synapse and neurotransmission genes in autism. Mol Psychiatry. 2012;17(4):402–11. doi:10.1038/mp.2011.10.

Gene Ontology C. Gene ontology consortium: going forward. Nucleic Acids Res. 2015;43(Database issue):D1049–56. doi:10.1093/nar/gku1179.

Genomeweb. With $999 whole genome sequencing service, Veritas embarks on goal to democratize DNA information [Press release]. 2016. Retrieved from https://www.genomeweb.com/sequencing-technology/999-whole-genome-sequencing-service-veritas-embarks-goal-democratize-dna?utm_source=SilverpopMailing&utm_medium=email&utm_campaign=Molecular%20Diagnostics%20Bulletin:%20With%20$999%20Whole-Genome%20Sequencing%20Service,%20Veritas%20Embarks%20on%20Goal%20to%20Democratize%20DNA%20Information%20-%2003/10/2016%2003:40:00%20PM.

Geschwind DH, State MW. Gene hunting in autism spectrum disorder: on the path to precision medicine. Lancet Neurol. 2015;14(11):1109–20. doi:10.1016/S1474-4422(15)00044-7.

Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott K M, Chenier S, . . . Brudno M. PhenoTips: patient phenotyping software for clinical and research use. Hum Mutat. 2013;34(8):1057–65. doi:10.1002/humu.22347.

Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, … e, MN. The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. Genet Med. 2013;15(10):761–71. doi:10.1038/gim.2013.72.

Green RC, Farahany NA. Regulation: the FDA is overcautious on consumer genomics. Nature. 2014;505(7483):286–7.

Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE, Glessner JT, … Polychronakos C. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature. 2007;448(7153):591–94. doi:10.1038/nature06010.

Hall MA, Verma SS, Wallace J, Lucas A, Berg RL, Connolly J, … Ritchie MD. Biology-driven gene-gene interaction analysis of age-related cataract in the eMERGE network. Genet Epidemiol. 2015;39(5):376–84. doi:10.1002/gepi.21902.

Hallmayer J, Cleveland S, Torres A, Phillips J, Cohen B, Torigoe T, … Risch N. Genetic heritability and shared environmental factors among twin pairs with autism. Arch Gen Psychiatry. 2011;68(11):1095–102. doi:10.1001/archgenpsychiatry.2011.76.

Hamosh A, Sobreira N, Hoover-Fong J, Sutton VR, Boehm C, Schiettecatte F, et al. PhenoDB: a new web-based tool for the collection, storage, and analysis of phenotypic features. Hum Mutat. 2013;34(4):566–71. doi:10.1002/humu.22283.

Healthcare IT News. Mayo Clinic takes aim at consumer genomics [press release]. 2015. Retrieved from: http://www.healthcareitnews.com/news/mayo-clinic-takes-aim-consumer-genomics

Knoppers B. Framework for responsible sharing of genomic and health-related data. HUGO J. 2014;8:6.

Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, … Robinson PN. The human phenotype ontology project: linking molecular biology and disease through phenotype data. Nucleic Acid. Res. 2014;42(Database issue):D966–74. doi:10.1093/nar/gkt1026.

Lawler M, Siu LL, Rehm HL, Chanock SJ, Alterovitz G, Burn J, … Health. All the world's a stage: facilitating discovery science and improved cancer care through the global alliance for genomics and health. Cancer Discov. 2015;5(11):1133–36. doi:10.1158/2159-8290.CD-15-0821.

Marsolo K, Spooner SA. Clinical genomics in the world of the electronic health record. Genet Med. 2013;15(10):786–91. doi:10.1038/gim.2013.88.

McGregor TL, Van Driest SL, Brothers KB, Bowton EA, Muglia LJ, Roden DM. Inclusion of pediatric samples in an opt-out biorepository linking DNA to de-identified medical records: pediatric BioVU. Clin Pharmacol Ther. 2013;93(2):204–11. doi:10.1038/clpt.2012.230.

Mungall CJ, Washington NL, Nguyen-Xuan J, Condit C, Smedley D, … Haendel MA. Use of model organisms and disease databases to support matchmaking for human disease gene discovery. Hum Mutat. 2015;36(10):979–84. doi:10.1002/humu.22857.

Namjou B, Keddache M, Marsolo K, Wagner M, Lingren T, Cobb B, … Harley JB. EMR-linked GWAS study: investigation of variation landscape of loci for body mass index in children. Front Genet. 2013;4:268. doi:10.3389/fgene.2013.00268.

Namjou B, Marsolo K, Lingren T, Ritchie MD, Verma SS, Cobb BL, … Harley JB. A GWAS study on liver function test using eMERGE network participants. PLoS One. 2015;10(9):e0138677. doi:10.1371/journal.pone.0138677.

Nishimura A, Carey J, Erwin PJ, Tilburt JC, Murad MH, McCormick JB. Improving understanding in the research informed consent process: a systematic review of 54 interventions tested in randomized control trials. BMC Med Ethics. 2013;14:28. doi:10.1186/1472-6939-14-28.

Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, … Rehm HL. The matchmaker exchange: a platform for rare disease gene discovery. Hum Mutat. 2015;36(10):915–21. doi:10.1002/humu.22858.

Rasmussen LV, Thompson WK, Pacheco JA, Kho AN, Carrell DS, Pathak J, … Starren JB. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. J Biomed Inform. 2014;51:280–86. doi:10.1016/j.jbi.2014.06.007.

Ridgeway JL, Han LC, Olson JE, Lackore KA, Koenig BA, Beebe TJ, Ziegenfuss JY. Potential bias in the bank: what distinguishes refusers, nonresponders and participants in a clinic-based biobank? Public Health Genomics. 2013;16(3):118–26. doi:10.1159/000349924.

Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, … Hakonarson H. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome Res. 2009;19(9):1682–90. doi:10.1101/gr.083501.108.

Shirts BH, Salama JS, Aronson SJ, Chung WK, Gray SW, Hindorff LA, … Overby CL. CSER and eMERGE: current and potential state of the display of genetic information in the electronic health record. J Am Med Inform Assoc. 2015;22(6):1231–42. doi:10.1093/jamia/ocv065.

Tarczy-Hornoch P, Amendola L, Aronson SJ, Garraway L, Gray S, Grundmeier RW, … Yang YA. survey of informatics approaches to whole-exome and whole-genome clinical reporting in the electronic health record. Genet Med. 2013;15(10):824–32. doi:10.1038/gim.2013.120.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, … DePristo MA. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11.10–11.33. doi:10.1002/0471250953.bi1110s43.

Vassy JL, Lautenbach DM, McLaughlin HM, Kong SW, Christensen KD, Krier J, … MedSeq P. The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. Trials. 2014;15:85. doi:10.1186/1745-6215-15-85.

Weitzel KW, Alexander M, Bernhardt BA, Calman N, Carey DJ, Cavallari LH, … Network I. The IGNITE network: a model for genomic medicine implementation and research. BMC Med Genomics. 2016;9(1):1. doi:10.1186/s12920-015-0162-5.

Yuen RK, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N, … Scherer SW. Whole-genome sequencing of quartet families with autism spectrum disorder. Nat Med. 2015;21(2):185–91. doi:10.1038/nm.3792.

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014;32(3):246–51. doi:10.1038/nbt.2835.

Chapter 16
Bioinformatics and Orphan Diseases

Anil G. Jegga

Abstract In general, a rare or orphan disease is any disease that affects a small percentage of the population. Since a majority of the known orphan diseases are genetic, they are present throughout the life of affected individuals. Many of the orphan diseases appear early in life and approximately 30 % of children with orphan diseases die before the age of 5. Further, a large majority of these diseases lack effective treatments. While most of genes and pathways underlying orphan diseases remain obscure, technological advances and innovative informatics approaches are expected to accelerate the rate of identification of underlying causal mutations and therapeutic discovery. Recent technological advances in DNA sequencing for instance, can aid in identifying genes associated with orphan diseases of previously unknown etiology using DNA from as few as 2–4 patients. Likewise, advanced computational statistical techniques permit integration and mining of omics data from orphan disease patients with high throughput "signatures" representing cellular responses to perturbing agents to identify therapeutic candidates for orphan diseases. In this chapter, we review some of the current bioinformatic analytical options available for orphan disease and drug research including computational approaches for candidate gene prioritization and high throughput compound screening to enable therapeutic discovery. We also discuss strategies and present examples and case studies of common drugs being repositioned for treatment of orphan diseases.

Keywords Rare disease • Orphan disease • Disease networks • Drug repositioning • Drug repurposing • Gene prioritization

A.G. Jegga, DVM, M.S. (✉)
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave. – MLC 7024, Cincinnati, OH 45229, USA

Department of Pediatrics, University of Cincinnati College of Medicine, 3333 Burnet Ave. – MLC 7024, Cincinnati, OH 45229, USA

Department of Computer Science, University of Cincinnati College of Engineering, 3333 Burnet Ave. – MLC 7024, Cincinnati, OH, USA
e-mail: Anil.Jegga@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_16

16.1 Introduction

A rare or orphan disease (OD) is any disease that affects a small percentage of the population. Most of the known ODs are genetic, and therefore are present throughout the life of an affected individual. Many appear early in life and about 30 % of children with ODs die before the age of five. In the United States, the Rare Disease Act of 2002 defines an OD as any disease or condition that affects fewer than 200,000 persons in the United States, while the European Commission on Public Health, defines ODs as those which are life-threatening or chronically debilitating and are of such low prevalence (1 in 2000 people) that special combined efforts are needed to address them. On the other hand, in Japan an OD is defined as one that affects fewer than 50,000 patients. While the incidence of an individual OD may be small, cumulatively, in the US itself, the 8000 known ODs affect about 25 million Americans, or nearly 10 % of the US population (Rados 2003). While some of the listed ODs are well-known and well-studied (e.g., cystic fibrosis), very little is known about a majority of ODs primarily because several ODs affect patient populations of fewer than a 100. Additionally, about 250 new ODs and conditions are described each year (Wastfelt et al. 2006).

Even though opportunities now exist to accelerate progress toward understanding the basis for many more ODs and for developing innovative medical approaches, relatively few efforts have successfully addressed scientific or technical questions across a spectrum of ODs (Field et al. 2010; Zhang et al. 2011). Constructing networks that underlie biological processes and pathways associated with ODs and identifying the functional units that respond to genetic perturbations and potentially affect disease risk or therapeutic response may facilitate systematic advancement of OD research and health care in a favorable direction (Zhang et al. 2011). For instance, studies of biological networks can identify common pathways or processes for multiple ODs that are biologically related and comprehensively understood. Such molecular basis may provide opportunities for interventions that are beneficial for an array of related ODs. Furthermore, this capability may open the door for the discovery of single therapies that can not only treat multiple ODs, but, potentially, also more common diseases (Field et al. 2010).

In this chapter, we review some of the current bioinformatics and genomics-based approaches to study the ODs and drug discovery process, including drug repositioning opportunities. We present several case studies and examples to illustrate these approaches. In the first section we outline some of the fundamental concepts and introduce organizations and resources related to orphan disease and drugs. In the second and third sections we specifically focus on bioinformatics and genomics centered approaches used to investigate ODs and discover drug targets and drug repositioning candidates for ODs.

16.2 The Orphan Drug Act, 1983

The United States' Orphan Drug Act (ODA; 1983) includes both rare diseases and any non-rare diseases for which there is no reasonable expectation that the cost of developing and making available a drug for such a disease in the US can potentially be recovered from sales of that drug in the US. Since the definitions of rare diseases refer to treatment availability, resource scarcity and disease severity, rare diseases are now commonly referred to as orphan diseases (ODs), especially after the orphan drug movement that began in the US in 1983. The ODA went into effect to encourage the development and marketing of drugs ("orphan drugs") to treat ODs and conditions.

The ODA evolved in response to the small number of orphan drugs that were approved in the US in the years prior to the approval of the ODA (United States Food and Drug 1992). About 8000 ODs have been identified, and a list is maintained by the Office of Rare Diseases (ORD) at the NIH. The Food and Drug Administration (FDA) administers the ODA and reviews applications for orphan designations. Some of the incentives provided to the drug companies involved in production of orphan drugs include 7-year marketing exclusivity to sponsors of approved orphan products, a tax credit of 50 % of the cost of conducting human clinical testing, and research grants for clinical testing of new therapies to treat ODs (ODA 2001). The ORD designates diseases that qualify and administers the small grants program, while the Center for Drug Evaluation and Research and the Center for Biologics Evaluation and Research review applications for marketing approval (ODA 2001).

16.3 Orphan Diseases and Drugs – Organizations and Resources

In the following sections, we briefly outline some of the organizations and resources focused and related to orphan diseases and drugs (see Table 16.1)

16.3.1 Genetic and Rare Diseases Information Center (GARD)

The Genetic and Rare Diseases Information Center was created in 2002 by the National Human Genome Research Institute (NHGRI) and the Office of Rare Diseases Research (ORDR) – two agencies at the National Institutes of Health (NIH) – to help people find useful information about genetic and ODs. In December 2011, NIH established the National Center for Advancing Translational Sciences (NCATS) to speed movement of discoveries from the laboratory to patients. NCATS now includes the ORDR, which coordinates and supports OD research, and TRND, the Therapeutics for Rare and Neglected Diseases, a program to encourage and speed the development of new drugs for rare and neglected diseases. With an aim to

Table 16.1 List of resources relevant to orphan diseases and drugs

Resource name	Description and URL
EURODIS	European organization for rare diseases; http://www.eurordis.org
List of marketed orphan drugs in Europe	http://www.fda.gov/orphan/designat/Approvals.htm
List of orphan drugs in Australia	http://www.tga.gov.au/docs/html/orphand2.htm
List of the American orphan diseases	http://rarediseases.info.nih.gov/RareDiseaseList.aspx
NIH rare diseases (GARD)	Genetic and Rare Diseases; http://rarediseases.info.nih.gov/GARD/
NORD	National Organization for Rare Disorders; http://www.rarediseases.org
OMIM	Online Mendelian Inheritance in Man; http://omim.org
OOPD at FDA	Office of Orphan Product Development; http://www.fda.gov/orphan
Orphan Disease Networks	Orphan disease and gene networks; http://research.cchmc.org/od
Orphan drugs at FDA	Orphan drug designations and approvals; http://www.accessdata.fda.gov/scripts/opdlisting/oopd/index.cfm
Orphanet	Portal for rare diseases and orphan drugs; http://www.orpha.net
RDRD from FDA	Rare disease Repurposing Database; http://www.fda.gov/ForIndustry/DevelopingProductsforRareDiseasesConditions/

work closely with partners in the regulatory, academic, nonprofit, and private sectors, NCATS goals are to identify and overcome hurdles that slow the development of effective treatments and cures.

16.3.2 European Organization for Rare Disorders (EURORDIS)

The European Organization for Rare Diseases, a European equivalent of the USA's GARD, is a patient-driven alliance of patient organizations and individuals active in the field of rare diseases. Its mission is to build a strong pan-European community of patient organizations and people living with rare diseases, to be their voice at the European level, and – directly or indirectly – to fight against the impact of rare diseases on their lives.

16.3.3 Orphanet

Orphanet (Ayme 2003) is the reference portal for information on rare diseases and orphan drugs. It is perhaps the best known and comprehensive database for ODs and drugs. It is led by a European consortium of around 40 countries, coordinated

by the French team. National teams are responsible for the collection of information on specialized clinics, medical laboratories, ongoing research and patient organizations in their country. The French coordinating team is responsible for the infrastructure of Orphanet, management tools, quality control, rare disease inventory, classifications and production of the encyclopedia. Orphanet is governed by various committees, which independently supervise the project in order to ensure its coherence, evolution and viability. However, because it predominantly follows European Union (EU) regulations and definitions of OD, there is a considerable difference between the US-based NIH list of rare diseases (http://rarediseases.info.nih.gov/RareDiseaseList.aspx) and the EU prevalence-based Orphanet list of rare diseases.

16.3.4 OMIM Database

The OMIM (Online Mendelian Inheritance in Man) database (Hamosh et al. 2000) is a comprehensive and authoritative compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources. OMIM while useful is not limited to rare conditions. Neither Orphanet nor OMIM databases have analytical servers that can facilitate further analysis (e.g., gene set enrichment analysis of all genes associated with a particular OD).

16.4 Mutant Gene Network Analyses of Orphan Diseases

Most of the OD-related research efforts focus on either a single OD or a small set of related ODs. Additionally, previous studies on disease networks did not separate out the ODs from the common diseases. To address these issues, we have conducted a bioinformatic-based global analysis of all ODs with known OD-causing mutant genes (ODMGs) to identify and investigate relationships based on shared genes or shared functional features (Zhang et al. 2011). Most importantly, using bioinformatics- and network analyses-based approaches, we tackled an interesting question of potential differences in properties between ODMGs and causal genes of common diseases. The orphan diseases and mutant gene information was downloaded from the Orphanet database (Ayme 2003). Starting with a bipartite network of known OD and OD-causing mutant genes (1772 ODs and 2124 ODMGs) and using the human protein interactome, we first constructed and topologically analyzed three networks (see Fig. 16.1): orphan disease network (ODN), orphan disease-causing mutant gene network (ODMGN) and orphan disease-causing mutant gene interactome (ODMGI).

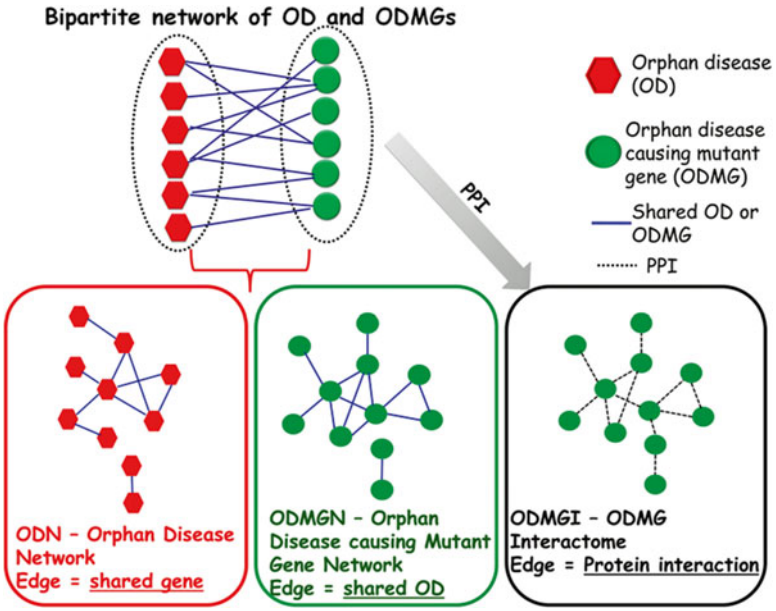


Fig. 16.1 Schematic representation of various networks related to ODs and ODMGs. Starting with a bipartite network (*top panel*) of orphan diseases (ODs; *red* hexagons) and their known causal genes (orphan disease-causing mutant genes or ODMGs; *green circles*), three types of networks are generated. The ODN (*lower panel left-hand side*) is an OD network where a node represents OD while the edge represents shared ODMG. The ODMGN (*lower panel, central pane*) represents an ODMG network with nodes being ODMGs while the edges represent shared ODs. The third network, ODMG interactome is built using the human protein-protein interactions and here the nodes are ODMGs while the edge represents a protein interaction between the ODMG-encoded proteins

In the ODN, nodes are ODs while edges are shared genes; the ODMGN comprised ODMGs as nodes and shared ODs as edges. The orphan disease gene interactome is the protein interactions network of ODMGs. We also compared these networks with earlier studies focusing on all diseases (Goh et al. 2007; Feldman et al. 2008) primarily to answer the question as to whether there is any difference in the network topology or functional properties of orphan disease genes and common disease genes. Network-based approaches helped us in finding that ODMGs are predominantly essential, a finding that is in contrast to the previous reports of disease genes being non-essential (Goh et al. 2007; Feldman et al. 2008). By integrating data from mouse knock-out models with the network-based approaches, we also found that ortholog gene knock-out models of orphan disease genes in the mouse in most cases either result in premature death or are embryonic lethals. Thus, informatics-based approaches integrating heterogeneous data sources (e.g., human orphan disease-causing mutant genes, protein interactome, and data from mouse knock-out models) and applying network analyses helped in understanding the orphan diseases better. For example, the two findings – ODMGs are predominantly

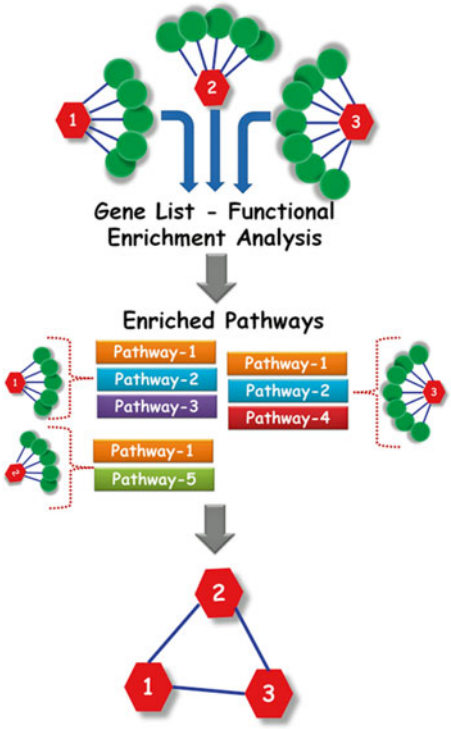


Fig. 16.2 Schematic representation of steps involved in building an orphan disease functional linkage network. ODs (e.g., 1, 2, 3; *red hexagons*) with known causal genes (*green circles*) are subjected to enrichment analyses (using ToppFun; http://toppgene.cchmc.org) to identify enriched features (e.g., pathways; colored rectangles). Two ODs are then connected if they share an enriched feature (pathway, in this case; blue edges represent shared enriched pathways). Thus, although ODs 1, 2 and 3 do not share any genes (*top panel*), they can still be connected via shared enriched pathways (*lower panel*).

essential and are associated with premature deaths – probably explain why most cases of orphan diseases are life-threatening. Most of the previous studies elucidating relationships between diseases are genecentric and therefore are limited in their discovery of new and unknown disease relationships (Suthram et al. 2010). To address this, we (Zhang et al. 2011) and others (Li and Agarwal 2009; Linghu et al. 2009; Suthram et al. 2010) recommend using functional linkage maps. As part of this study, we therefore built functional linkage networks (Fig. 16.2) of ODs based on shared significant pathways or biological processes rather than just shared genes.

Briefly, we first extracted all those ODs with four or more mutant genes from our original data set. Starting with this filtered sub-bipartite network of 196 ODs and 1087 genes (1283 total nodes and 1395 total edges), we built OD-OD networks based on shared genes and shared functions. The enriched functions ($p<0.05$) for each of the 196 ODs were determined with the ToppFun application (http://toppgene.cchmc.org) (Chen et al. 2009b). Using the enriched features for each of the orphan diseases, we rebuilt the orphan disease networks. However, this time the edge between two ODs represents an enriched shared function (BP, CC, Pathway, or MP) and not necessarily a shared gene (Fig. 16.2). After generating these function-based OD networks, we compared them with the gene-based orphan disease networks to find the overlapping nodes and edges. The results were surprising. Although the node agreement between the gene-based and function-based ODNs

was relatively higher, the edge agreement was much lower and indicates that their wiring is significantly different. This suggests that the relationship between the ODs cannot be fully captured by the gene-based networks alone. Thus, by considering functional connectivity between causative genes involved in different orphan diseases, relationships between orphan diseases that are based on underlying molecular mechanisms can be revealed. Such associations can potentially be used to generate novel hypotheses on the molecular mechanisms of diseases, and can in turn guide the development of relevant therapy (Linghu et al. 2009) or potential drug repositioning candidates (Zhang et al. 2011).

Apart from leading to new insights into the biological underpinnings of various orphan diseases, a global analysis of *orphan diseasome* will encourage the development of new and innovative research on these rare conditions, which have been hitherto understudied. Additionally, the global analysis of all ODs can help in analyzing co-morbidities and the underlying molecular basis apart from establishing potential networking opportunities. The functional linkage networks of ODs apart from the conventional gene-based connectivity maps of diseases have direct implications to drug discovery process. For more details the readers are referred to the original publication (Zhang et al. 2011).

16.5 Orphan Disease Gene Identification and Prioritization – Computational Approaches

Despite the advances in genomewide techniques such as linkage analysis and association studies, the selected disease loci are actually a region of the chromosome rather than location of a "gene". A "locus" identified in this way may contain several hundreds of candidate genes. For instance, in the OMIM database, over 900 ODs are described that have been mapped to one or more such gene map loci and are classified as having an 'unknown molecular basis' (OMIM IDs prefixed with "#"). The prioritization of the positional candidate genes in these OD loci is therefore an important step to facilitate OD-gene identification for further experimental analysis.

As shown in Table 16.2, several candidate gene prioritization methods have been developed to overcome the limitations of high-throughput, genome-wide studies like linkage analysis and gene expression profiling, both of which typically result in the identification of hundreds of potential candidate genes (Freudenberg and Propping 2002; Turner et al. 2003; Adie et al. 2005, 2006; Tiffin et al. 2005, 2006; Aerts et al. 2006; Chen et al. 2007, 2009b; Thornblad et al. 2007; Zhu and Zhao 2007) (for additional details see (Piro and Di Cunto 2012)). Most of these computational approaches are based on the assumption that similar phenotypes are caused by genes with similar or related functions (Jimenez-Sanchez et al. 2001; Smith and Eyre-Walker 2003; Turner et al. 2003; Goh et al. 2007; Chen et al. 2009b). They, however, differ by the data sources utilized and by the strategy adopted in calculat-

Table 16.2 List of current bioinformatics approaches and tools to rank human disease candidate genes

Approach	Online availability	Data types used	Training set (Input)
Approaches based on disease gene properties			
DGP (Lopez-Bigas and Ouzounis 2004)	http://cgg.ebi.ac.uk/services/dgp/	Sequence	Not applicable (N/A)
PROSPECTR (Adie et al. 2005)	http://www.genetics.med.ed.ac.uk/prospectr/	Sequence	N/A
Approaches using links between genes and phenotypes			
Genes2Diseases (Perez-Iratxeta et al. 2002, 2005)	http://www.ogic.ca/projects/g2d_2/	Sequence, Gene Ontology (GO), literature mining	Phenotype GO terms Known genes
BITOLA (Hristovski et al. 2005)	http://www.mf.uni-lj.si/bitola/	Literature mining	Concept
GeneSeeker (van Driel et al. 2003, 2005)	http://www.cmbi.ru.nl/GeneSeeker/	Expression, phenotype, literature mining	N/A
GFINDer (Masseroli et al. 2004, 2005)	http://www.bioinformatics.polimi.it/GFINDer/	Expression, phenotype	N/A
TOM (Rossi et al. 2006)	http://www-micrel.deis.unibo.it/~tom/	Expression, GO	Known genes and/or disease loci
Approaches using functional relatedness between candidate genes			
OMIM phenome map (van Driel et al. 2006)	http://www.cmbi.ru.nl/MimMiner/	Phenotype, sequence, GO, protein interactions	N/A
SUSPECTS (Adie et al. 2006)	http://www.genetics.med.ed.ac.uk/suspects/	Sequence, expression, GO	Known genes
Prioritizer (Franke et al. 2006)	http://www.prioritizer.nl/	Expression, GO, protein interactions	Disease loci
Endeavour (Aerts et al. 2006)	http://www.esat.kuleuven.be/endeavour/	Sequence, expression, GO, pathways, literature mining	Known genes
ToppGene (Chen et al. 2009b)	http://toppgene.cchmc.org	Mouse phenotype, expression, GO, pathways, literature mining	Known genes
ToppNet (Chen et al. 2009a)	http://toppgene.cchmc.org	Protein interactions	Known genes

The first column has the source or the name of the tool (including reference, if available) while the second column has the URL of the corresponding web application. The third column shows the list of genomic annotation types/features used by each of the methods for candidate gene ranking. The last column has details of the training or the input data, if used (Note: modified from Kaimal et al. (2011), this list is extensive, but not exhaustive.; reference (Piro and Di Cunto 2012) provides an additional list of tools)

ing similarity (Tranchevent et al. 2008). Except for ENDEAVOUR (Aerts et al. 2006; Tranchevent et al. 2008) and ToppGene (Chen et al. 2007, 2009b), most of the existing approaches utilize a very limited number of data sources.

Because biological networks have been found to be comparable to communication and social networks (Junker et al. 2006) through commonalities such as scale-freeness and small-world properties, the algorithms used to analyze social and Web networks have been successfully used for disease gene identification and ranking. These network-based approaches predominantly use protein-protein interaction networks (PPIN) and the candidate genes are typically ranked based on their connectivity to known disease genes (seed or training set). While PPINs have been used widely to identify novel disease candidate genes (George et al. 2006; Xu and Li 2006; Kann 2007; Kohler et al. 2008; Wu et al. 2008), several recent studies (Chen et al. 2006, 2009a; Kohler et al. 2008; Wu et al. 2008; Ortutay and Vihinen 2009) report also using them for candidate gene prioritization.

A principal limitation of almost all of the current disease gene identification and prioritization approaches is that they are gene-centric. However, it has been speculated that complex traits result more often from noncoding regulatory variants than from coding sequence variants (King and Wilson 1975; Mackay 2001; Korstanje and Paigen 2002), yet analyzing noncoding regulatory variants is replete with problems. For instance, functional consequences of coding region variants are relatively easy to assess (e.g., missense, nonsense, splicing, etc.) while interpreting the consequences of noncoding sequence variants is more complicated and the relationships between noncoding sequence variation and gene expression level or phenotypes are relatively less well understood (Kaimal et al. 2011).

16.6 Exome Sequencing to Decipher Orphan Diseases

Next generation whole genome or exome sequencing can be used to search for Mendelian disease genes in an unbiased manner (Gilissen et al. 2012). Of the ODs with a known causal gene mutation, about 70 % are monogenic (Zhang et al. 2011) and as per the current version of OMIM, there are about 5000 monogenic ODs and for half of these the underlying genes remain unknown. The OD causal gene identification thus represents the first step to a better understanding of the pathophysiological mechanisms underlying ODs, which in turn can lead to developing effective therapeutic interventions. While massively parallel DNA sequencing technologies have rendered the whole genome resequencing of individual humans increasingly practical, the associated expenses continue to be a major hurdle. However, about 85 % of the known genetic causes for Mendelian disorders affect the protein coding exonic regions (Botstein and Risch 2003), which accounts for approximately 2 % of the genome. An alternative approach involves the targeted re-sequencing of all protein-coding subsequences (exome sequencing), potentially overcoming the financial hurdle to routine comprehensive sequencing for diagnostic purposes in the near future (Bainbridge et al. 2011; Kingsmore and Saunders 2011). Thus, whole

exome sequencing using the next generation technologies provides a new and transformational approach for identifying causative mutations in ODs (see Table 16.3) and is likely to become the most commonly used tool for disease gene identification (Gilissen et al. 2012).

Although there is no doubt about the promise of exome sequencing in OD research, it has some limitations. First, the cause of an OD could be a noncoding variation or a large indel or structural genomic variant, all of which are missed by exomic sequencing. Second, some of the variants may not be identified because of lack of sequence coverage of the variant or could be related to technical errors (e.g.

Table 16.3 Orphan disease gene identification by exome sequencing

Orphan disease	Identified OD-Gene	References
Ablepharon Macrostomia and Barber-Say Syndromes	*TWIST2*	Marchegiani et al. (2015)
Adams-Oliver syndrome	*DLL4*	Meester et al. (2015)
Amelogenesis imperfecta	*FAM20A*	O'Sullivan et al. (2011)
Amyotrophic lateral sclerosis	*VCP*	Johnson et al. (2010)
Autoimmune lymphoproliferative syndrome	*FADD*	Bolze et al. (2010)
Cerebral cortical malformations	*WDR62*	Bilguvar et al. (2010)
Chondrodysplasia and abnormal joint development	*IMPAD1*	Vissers et al. (2011)
Combined hypolipidemia	*ANGPTL3*	Musunuru et al. (2010)
Congenital chloride diarrhea	*SLC26A3*	Choi et al. (2009)
Fowler syndrome	*FLVCR2*	Lalonde et al. (2010)
Hajdu-Cheney syndrome	*NOTCH2*	Simpson et al. (2011), Majewski et al. (2011), and Isidor et al. (2011)
Heimler syndrome	*PEX1, PEX6*	Ratbi et al. (2015)
Hereditary spastic paresis	*KIF1A*	Erlich et al. (2011)
Hyperphosphatasia mental retardation syndrome	*PIGV*	Krawitz et al. (2010)
Infantile mitochondrial cardiomyopathy	*AARS2*	Götz et al. (2011)
Kabuki syndrome	*MLL2*	Ng et al. (2010)
Kaposi sarcoma	*STIM1*	Byun et al. (2010)
Miller syndrome	*DHODH*	Ng et al. and Antonarakis and Beckmann (2006)
Olmsted syndrome	*TRPV3*	Lin et al. (2012)
Osteogenesis imperfecta	*SERPINF1*	Becker et al. *and* Choi et al. (2009)
Perrault syndrome	*HSD17B4*	Pierce et al. (2010)
Progeroid syndrome	*BANF1*	Puente et al. (2011)
Retinitis pigmentosa	*DHDDS*	Züchner et al. (2011)
Schinzel-Giedion syndrome	*SETBP1*	Hoischen et al. (2010)
Sensenbrenner syndrome	*WDR35*	Gilissen et al. (2010)
Spinocerebellar ataxia	*TGM6*	Wang et al. (2010)

bioinformatics variant calling issues) (Gilissen et al. 2012). While the technology and software for sequencing are widely and readily available, there is an increasing need for better and novel computational approaches to annotate the variants, integrate and mine the variant data, prioritize the variants and candidate genes, and analyze pathways for potential genetic heterogeneity. In the following sections, we first review a few of the recent studies that evaluated available bioinformatic applications for their predictive performance of causal variant identification and analysis. Following this, we will present a few case studies from recently published studies, where disease-network analysis approach is used along with exome sequencing to identify novel OD causal variants.

16.7 Bioinformatic Approaches for Variant Prioritization

A recent comparative study (Dong et al. 2015) evaluated several bioinformatics-based approaches for variant prioritization. Dong et al. (2015) evaluated quantitatively and qualitatively the predictive performance of 18 existing variant deleteriousness prediction methods. These included eleven function prediction-based scoring (SIFT (Kumar et al. 2009), PolyPhen (Adzhubei et al. 2010), MutationTaster (Schwarz et al. 2010), PANTHER (Thomas et al. 2003), PhD-SNP (Capriotti et al. 2006), SNAP (Bromberg and Rost 2007), SNPs&GO (Calabrese et al. 2009), MutPred (Li et al. 2009), FATHMM (Shihab et al. 2013), Mutation Assessor (Reva et al. 2011), and LRT (Chun and Fay 2009)), three conservation-based scoring (GERP++ (Davydov et al. 2010), SiPhy (Garber et al. 2009), and PhyloP (Cooper et al. 2005)), and four ensemble scoring (CADD (Kircher et al. 2014), PON-P (Olatubosun et al. 2012), KGGSeq (Li et al. 2012), and CONDEL (Gonzalez-Perez and Lopez-Bigas 2011)) methods (Table 16.4).

To facilitate more accurate variant prediction, the authors developed and evaluated two ensemble-based approaches, namely, Support Vector Machine (SVM) and Logistic Regression (LR) based models. To compare the performance, the authors manually collected four datasets (one for training SVM and LR model and three for testing) of nonsynonymous single nucleotide variants (nsSNVs) based on the Uniprot database (UniProt 2014, 2015), CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) sequencing project (ARIC 1989; Morrison et al. 2013), VariBench dataset (Sasidharan Nair and Vihinen 2013), and nsSNVs from studies published in the journal Nature Genetics. Of the 18 tools evaluated, the authors found that FATHMM and KGGSeq had the highest discriminative power among independent scores and ensemble scores, respectively. To demonstrate the value of combining information from multiple prediction algorithms, the authors developed two new ensemble scores that integrate nine independent scores and allele frequency. These ensemble scores were found to have the highest discriminative power in comparison to all other deleteriousness prediction scores tested. They were also shown to have relatively low false-positive prediction rate for benign yet rare nonsynonymous variants (Dong et al. 2015).

Table 16.4 Bioinformatic resources for variant annotation

Resource	URL
1000 Genomes Project	http://www.1000genomes.org/
ANNOVAR	http://www.openbioinformatics.org/annovar/
CADD	http://cadd.gs.washington.edu/
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/
CONDEL	http://bg.upf.edu/condel/home
dbNSFP	http://sites.google.com/site/jpopgen/dbNSFP
GERP++	http://mendel.stanford.edu/SidowLab/downloads/gerp/
Exome Variant Server	http://evs.gs.washington.edu/EVS/
Exomiser	http://www.sanger.ac.uk/science/tools/exomiser
eXtasy	http://extasy.esat.kuleuven.be/
KGGSeq	http://statgenpro.psychiatry.hku.hk/limx/kggseq/
LRT	http://www.genetics.wustl.edu/jflab/lrt_query.html
MutationTaster	http://www.mutationtaster.org/
MutPred	http://mutpred.mutdb.org/
PANTHER	http://www.pantherdb.org/tools/csnpScoreForm.jsp
PhD-SNP	http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi
Phevor2	http://www.yandell-lab.org/software/phevor.html
PhyloP	http://compgen.bscb.cornell.edu/phast/; http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phyloP44way/
PolyPhen-2	http://genetics.bwh.harvard.edu/pph2/
PON-P	http://bioinf.uta.fi/PON-P/
SIFT	http://sift.jcvi.org/
SNAP	http://rostlab.org/services/snap/
SNPs&GO	http://snps-and-go.biocomp.unibo.it/snps-and-go/
Uniprot	http://www.uniprot.org/
VAAST 2	http://www.yandell-lab.org/software/vaast.html
Variant effect predictor	http://useast.ensembl.org/Homo_sapiens/Tools/VEP

16.8 Exome Sequencing and Bioinformatics Applications to Identify Novel Orphan Disease Causal Variants – Two Case Studies

In a recent study, Erlich et al. (2011) demonstrated how gene prediction tools, when used in combination with traditional mapping approaches, can be successfully applied to prioritization of OD candidate genes from exome resequencing experiments. This study illustrates the potential of combining genomic variant and gene level information to identify and rank novel causal variants of orphan diseases. The authors used three different candidate gene prioritization tools (Endeavour (Aerts et al. 2006), ToppGene (Chen et al. 2009b), and SUSPECTS (Adie et al. 2006)) to

prioritize *KIF1A* as the most likely candidate gene for hereditary spastic paraparesis (HSP). In this study, a familial case of hereditary spastic paraparesis (HSP) was analyzed through whole-exome sequencing and the four largest homozygous regions (containing 44 genes) were identified as potential HSP loci. The authors then applied several filters to narrow down the list further. For example, a gene was considered as potentially causative if it contains at least one variant that is either under purifying selection or not inherited from the parents or absent in dbSNP or the 1000 Genomes Project data. Because the majority of the known orphan disease variants affect coding sequences, they also checked whether the variant is non-synonymous. After this filtering step, 15 candidate genes were identified and this list was further prioritized using three computational methods (Endeavour (Aerts et al. 2006), ToppGene (Chen et al. 2009b), and SUSPECTS (Adie et al. 2006)). As a training set, a list of 11 seed genes associated with a pure type of HSP was compiled through literature mining. Interestingly, the top-ranking gene from all the three bioinformatics approaches (each of which uses different types of data and algorithms for prioritization) was *KIF1A*. Subsequent Sanger sequencing confirmed that *KIF1A* indeed is the causative variant.

In a second study, Benitez et al. (2011) used disease-network analysis approach as supporting in silico evidence of the role of the adult neuronal ceroid-lipofuscinosis (NCL) candidate genes identified by exome sequencing. The authors used Endeavour (Aerts et al. 2006) and ToppGene (Chen et al. 2009b) to rank the NCL candidate variant genes identified by exome sequencing. Known causal genes of other NCLs along with genes that are associated with phenotypically close disorders were used as training set for both the softwares (ToppGene and Endeavour). Interestingly, the three variants identified by exome sequencing (*PDCD6IP, DNAJC5* and *LIPJ*) were in the top five genes in the combined analysis using ToppGene and Endeavour suggesting that they may be functionally or structurally related with NCLs encoded genes and constituting true causative variants for adult NCL.

16.9 Drug Repositioning Strategies

Drug development in general is complicated, time-consuming, and expensive with extremely low success rates. To overcome or by-pass this productivity gap, more and more companies are resorting to "*Drug Repositioning*" or "*Drug Repurposing,*" or simply identifying and developing new uses for existing or abandoned pharmacotherapies (Ashburn and Thor 2004). This approach can significantly reduce the risks associated with drug development. Repositioned drugs can enter clinical phases more rapidly and at a lower cost than novel compounds because the starting point is usually approved compounds with known bioavailability and safety profiles, proven formulation and manufacturing routes, and well-characterized pharmacology (Boguski et al. 2009). It is therefore no surprise that in recent years, of the new medicines that reach their first markets, repurposed drugs account for ~30 %!

Following the FDA Orphan Drug Act in 1983, there has been a dramatic rise in new treatment options for ODs (~325 orphan drugs now available in the market).

However, these drugs cover only about 5 % of known ODs. Bypassing the traditional drug discovery process and discovering OD indications for already approved compounds could therefore be a viable strategy to jump-start the orphan drug discovery process. The recent release of a 235-drug database (Xu and Cote 2011) of approved compounds and products that show promise in ODs by the FDA further supports the need for a more systematized analysis of approved drugs for novel OD indications.

Drug repositioning is primarily based on two key principles: (a) the "promiscuous" nature of the drug and (b) targets relevant to a specific disease or pathway may also be critical for other diseases or pathways (Pujol et al. 2010; Sardana et al. 2011). While understanding the potential off-target interactions of existing drugs is of major interest in pharmaceutical research to understand molecular basis of adverse drug reactions and for discovering novel indications of drugs, identifying these targets in a context that also provides basic information on medical exploitability is a major challenge. In this context, recent rapid advances in genomic, proteomic, functional, and systems studies of the known drug targets and disease proteins have enabled the discovery of drugs, multitarget agents, combination therapies, and analysis of on-target and off-target toxicity and pharmacogenetic responses (Padhy and Gupta 2011). Such information is now publicly available in the form of different databases (e.g., DrugBank (Wishart et al. 2006), PharmGKB (Altman 2007), STITCH (Kuhn et al. 2012), Therapeutic Target Database (Zhu et al. 2012b)) which provide target and drug profiles. Druggable targets on the other hand can be identified integrating biochemistry and cell biology with genetics and physiology, as well as bioinformatics and network analysis-based approaches. Potential candidates can also be identified using FDA's "Disc" (Discontinued Drug Products) list, which contains thousands of drugs that made it through Phase I testing, but were withdrawn subsequently for reasons other than safety.

Although there are several advantages, rational drug repositioning poses formidable challenges, especially in the case of ODs because the molecular basis and the underlying mechanism of most ODs and drug actions are unknown, intricate, or are not amenable to human or computational data mining techniques. In the following sections we will discuss some of these issues and hurdles and present an overview of current strategies to overcome them to translate pharmacological and biomedical discoveries into safe and effective OD therapeutics (Sardana et al. 2011).

With the value of drug repositioning becoming increasingly evident, a number of companies have developed systematic approaches which can be broadly categorized into three strategies (see recent review by Sardana et al. (2011)) which are presented below along with case studies

16.9.1 Strategy 1: Knowledge-Based Drug Repositioning

This strategy involves integration of the accumulating heterogeneous pharmacological, genomic, biomedical, and chemical data and mining it, using novel analytical algorithms and approaches. The "virtual screening" is performed to discover

unrecognized or non-explicit connections between a drug, target, and disease. For example, Iorio et al. (2010a) developed MANTRA (Mode of Action by Network Analysis), a bioinformatics tool to analyze novel drugs and potential drug repositioning candidates of known and FDA approved drugs by the assignment of previously unrecognized putative therapeutic applications. The analysis is based on a novel similarity measure among the cellular responses elicited by a large set of compounds in humans. For each compound, this response is summarized by a prototype genome-wide ranked list (PRL) of genes. This PRL is built *in silico* by combining genome-wide profiles of differential expression, following treatments with the compound on a number of different human cell lines. The PRLs are then compared to each other with a novel approach based on Gene Set Enrichment Analysis (GSEA) ending up with an exhaustive set of drug pair-wise distance values. Using these distance measures, a drug-drug network is built and analyzed to identify modules or highly interconnected subnetworks or communities. Communities enriched for drugs with similar mechanism of action are then identified using known knowledge and literature searches. These communities are used to make novel hypotheses on known drugs and to classify novel drugs, when they are integrated in the network, hence the rays of light radiating from the man to the painted elements. Using this framework, the authors discovered that fasudil (a Rho-kinase inhibitor) could potentially be repositioned as treatment for neurodegenerative disorders.

In another study, Suthram et al. (2010) integrated molecular profiles of diseases with protein–protein interaction data, to infer protein functional modules and networks that were shared among many diseases. The molecular profiles of diseases were obtained from the NCBI Gene Expression Omnibus (GEO) (Barrett et al. 2007), while the protein-protein interaction (PPI) data for human was obtained from the Human Protein Reference Database (HPRD) (Goel et al. 2012). Using this human PPI data that was organized into 'modules' of functionally interacting proteins, a statistical approach was used to evaluate the molecular signatures of diseases for gene functional module activity, whereby the module activity was determined as the mean normalized transcriptional activity of its component genes in the disease molecular profile. A disease-disease network was then formed on the basis of functional module activity shared between diseases and mined to identify subnetworks with multiple known drug targets.

In another interesting study, Chiang and Butte (2009) used a "guilt by association" approach to discover alternative uses for drugs by systematically evaluating a drug treatment-based view of diseases. The authors used the DRUGDEX System, as a gold-standard comprehensive pharmacopeia resource. The DRUGDEX system provides literature-backed drug information on both (FDA) approved indications and off-label uses and is an U.S. government approved source for medical insurance reimbursable off-label uses. Using this system, the authors systematically captured the FDA-approved (FDA approved drug and indications) and practiced (FDA-approved drugs, but off-label drug uses) views into a Drug-Disease Knowledge Base consisting of 726 diseases and 2022 drugs. Given that the treatment profiles between a small subset of disease pairs can be quite different between the FDA-approved and practiced views, the authors reasoned that these discrepancies can be

mined to suggest novel drug uses. This was based on the hypothesis that if two diseases share similar therapies, then drugs that are currently used for only one of the two diseases may also be extrapolated to the other.

16.9.2 Strategy 2: Rescreening the Pharmacopeia Against New Targets

Taking advantage of the recent and rapid advances in high throughput screening technology, this unbiased strategy takes the semi-blind approach of re-screening existing compounds against a multitude of targets to identify possible therapeutic benefits or side-effects (O'Connor and Roth 2005). For instance, focusing on neglected parasitic diseases, researchers at the University of Dundee carried out a large scale screening to cover a broad range of potential targets (Brenk et al. 2008). Out of 2.3 million commercially available compounds about 222,000 compounds were selected for an in silico library, ~57,000 for a diverse general screening library, and ~1700 compounds for a focused kinase set. To compile these libraries, the authors set several rules to define unwanted groups and also to identify "lead-like" compounds which facilitate straightforward structure–activity relationship exploration. To assemble the focused library for hit discovery for kinases, literature and patents were mined and reviewed. Screening the known drugs (common and orphan) against OD causative and druggable proteins (Russ and Lampel 2005) can not only identify potentially new therapeutic agents for validated targets, but also identify and validate new orphan disease-relevant targets and potential drug combinatorials (Sardana et al. 2011). Apart from jump-starting orphan disease therapeutic programs, if a novel target is discovered for an existing drug (common or orphan), its chemical scaffold can be used as a good starting point to identify potential new chemical entities with the premise of developing them as drugs for the novel target (Grau et al. 2005).

16.9.3 Strategy 3: End-Point Screening

The starting point in this strategy is a phenotype of interest and the existing drugs are screened to discover drugs that produce an unanticipated (either as on- and/or off-target effects of compounds), yet desired, phenotypic result. For example, increasing autophagy may provide clinical benefit in the treatment of various diseases, and therefore there is a great effort in developing drugs enhancing this function (Iorio et al. 2010b). Using autophagy as a phenotype of interest, Iorio et al. (2010b) used a known inducer of autophagy (2-deoxy-D-glucose) and generated a list of drugs that were predicted to share a similar mode of action and identified a novel autophagy enhancer drug (fasudil) along with other previously known

inducers of autophagy. The authors used the Connectivity Map (Lamb et al. 2006) dataset which is a large compendia of publicly available gene expression data obtained by treatments of several human cell lines with a large collection of small molecules. A drug network is then built considering similarities in the consensual transcriptional responses and groups of highly interconnected nodes (drugs or small molecules) are then identified. These communities of highly interconnected nodes are subjected to further computational analyses to identify their enriched modes of action.

16.10 Big Data Integration and Mining for Drug Discovery and Drug Repositioning

The biomedical BIG data, in which biomolecular structural, functional and process knowledge is embedded consists of a large number of both structured relational databases and unstructured free-text publications. Migration from such information silos towards knowledge is facilitated by establishing higher order connectivity among the subsets taken from multiple domains. For example, a module consisting of a group of genes, pathways, diseases, drugs, and a group of drug-related adverse events (AEs) forms a meaningful multi-domain module, especially when various individual databases attest to dependences among pairs of subsets contained in the module. This larger module of apparently interrelated genes, drugs, pathways, and AEs takes us closer to answering the how question about the underlying phenomena. And a better answer to the how question will help generate better drug repositioning hypotheses.

Notable computational approaches and efforts have been made in connecting the biological and chemical domains (e.g. KEGG and DrugBank (Kanehisa et al. 2006; Wishart et al. 2008)). However, these databases were not designed for connecting with phenotypic information. While PharmGKB offers access to phenotypic information about several drugs along with their associated genotypes, it is not designed to represent the mechanism of actions of these therapeutic drugs (Altman 2007). Lamb *et al.* reported a promising "connectivity map" approach which associates small molecules, genes, and diseases through genomic profiling connections (Lamb et al. 2006; Lamb 2007). In another landmark study, Bork and colleagues used side-effect similarities to infer whether two drugs share a target (Campillos et al. 2008). In two seminal studies Keiser et al. and Kinnings et al. reported a structural similarity based chemoinformatics approach that identify the targets responsible for the polypharmacology of known drugs and potential new indications (Keiser et al. 2009; Kinnings et al. 2009). Although these approaches mark a great advancement in linking drugs to underlying diseases and drug repositioning, since these functional connections are purely based on one or two dimensional data (i.e., gene expression profiling or structural similarity), the insufficient coverage and limited

diversity of the perturbagens in the data collection continues to be a major shortcoming. The Library of Integrated Cellular Signatures (LINCS – http://www.lincsproject.org), a NIH-funded program is one of the efforts to overcome the shortcoming related to insufficient coverage. The LINCS datasets, for instance, include gene expression signatures for about 20K small-molecule compounds selected from sources such as known drugs, pathway-specific tool compounds, and compounds identified in NIH-sponsored small-molecule screening efforts. Researchers can now query disease-specific gene expression signatures against the 1.4 million small molecule expression profiles available in the LINCS database to identify potential candidate therapeutics. However, these gene expression-based therapeutic discoveries should be complemented with additional information including prior knowledge (e.g., pathways, biological processes, biological networks, etc.) to formulate mechanism of action related hypotheses. In our previous studies with rare disease networks (Zhang et al. 2011; Zhu et al. 2012a) and network based approaches for drug repositioning candidate discovery (Wu et al. 2013), we noted that the relationship between diseases or between diseases and drugs cannot be fully captured by the genes network alone.

16.11 FDA's Rare Disease Repurposing Database (RDRD)

Recently, the U.S. FDA's Office of Orphan Products Development has established a new resource: Rare Disease Repurposing Database (RDRD) (Xu and Cote 2011) for drug developers. It is a compilation of drugs that have shown promise for treating orphan diseases and already have FDA approval or designation. It is a database of products that (a) have received orphan status designation (i.e., they have been found "promising" for treating a rare disease); and (b) are already market-approved for the treatment of some other diseases. Since these compounds already have FDA approval, it is anticipated that repositioning these drugs for a new orphan disease indication will be quick and inexpensive for the developer, and will therefore help patients by getting to market quicker. The data are provided as three downloadable excel files: (a) Orphan-designated products with at least one marketing approval for a common disease indication; (b) Orphan-designated products with at least one marketing approval for an OD indication; and (c) Orphan-designated products with marketing approvals for both common and OD indication. These lists offer sponsors a new tool and a "shorter path" for finding special opportunities to develop niche therapies instead of beginning with an untested new therapy compound and obtaining an FDA approval. Additionally, these data sets can be used as a "positive control" or "gold standard" for testing or validating high-throughput and computational approaches for drug repositioning for orphan diseases.

References

Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinf. 2005;6:55.

Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics. 2006;22:773–4.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.

Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, et al. Gene prioritization through genomic data fusion. Nat Biotechnol. 2006;24:537–44.

Altman RB. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. Nat Genet. 2007;39:426.

Antonarakis SE, Beckmann JS. Mendelian disorders deserve more attention. Nat Rev. 2006;7:277–82.

ARIC. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. Am J Epidemiol. 1989;129:687–702.

Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004;3:673–83.

Ayme S. [Orphanet, an information site on rare diseases]. Soins. 2003;46–47.

Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, Newsham I, Reid JG, Fink JK, Morgan MB, Gingras MC, et al. Whole-genome sequencing for optimized patient management. Sci Transl Med. 2011;3:87re83.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res. 2007;35:D760–5.

Benitez BA, Alvarado D, Cai Y, Mayo K, Chakraverty S, Norton J, Morris JC, Sands MS, Goate A, Cruchaga C. Exome-sequencing confirms DNAJC5 mutations as cause of adult neuronal ceroid-lipofuscinosis. PLoS One. 2011;6:e26741.

Bilguvar K, Ozturk AK, Louvi A, Kwan KY, Choi M, Tatli B, Yalnizoglu D, Tuysuz B, Caglayan AO, Gokben S, et al. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature. 2010;467:207–10.

Boguski MS, Mandl KD, Sukhatme VP. Drug discovery. Repurposing with a difference. Science (New York, NY). 2009;324:1394–5.

Bolze A, Byun M, McDonald D, Morgan NV, Abhyankar A, Premkumar L, Puel A, Bacon CM, Rieux-Laucat F, Pang K, et al. Whole-exome-sequencing-based discovery of human FADD deficiency. Am J Hum Genet. 2010;87:873–81.

Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33(Suppl):228–37.

Brenk R, Schipani A, James D, Krasowski A, Gilbert IH, Frearson J, Wyatt PG. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. Chem Med Chem. 2008;3:435–44.

Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35:3823–35.

Byun M, Abhyankar A, Lelarge V, Plancoulaine S, Palanduz A, Telhan L, Boisson B, Picard C, Dewell S, Zhao C, et al. Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. J Exp Med. 2010;207:2307–12.

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat. 2009;30:1237–44.

Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008;321:263–6.
Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006;22:2729–34.
Chen JY, Shen C, Sivachenko AY. Mining Alzheimer disease relevant proteins from integrated protein interactome data. Pac Symp Biocomput. 2006:367–378.
Chen J, Xu H, Aronow BJ, Jegga AG. Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinf. 2007;8:392.
Chen J, Aronow BJ, Jegga AG. Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinf. 2009a;10:73.
Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009b;37:W305–11.
Chiang AP, Butte AJ. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. Clin Pharmacol Ther. 2009;86:507–10.
Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009;106:19096–101.
Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553–61.
Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15:901–13.
Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6:e1001025.
Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24:2125–37.
Erlich Y, Edvardson S, Hodges E, Zenvirt S, Thekkat P, Shaag A, Dor T, Hannon GJ, Elpeleg O. Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. Genome Res. 2011;21:658–64.
Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A. 2008;105:4323–8.
Field MJ, Boat TF. Rare diseases and orphan products: accelerating research and development. In: MJ Field, TF Boat, editors. Rare diseases and orphan products: accelerating research and development. Committee on Accelerating Rare Diseases Research and Orphan Product Development, Institute of Medicine. 2010.
Franke L, Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet. 2006;78:1011–25.
Freudenberg J, Propping P. A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics. 2002;18 Suppl 2:S110–5.
Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25:i54–62.
George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA. Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucleic Acids Res. 2006;34:e130.
Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van Lier B, Steehouwer M, van Reeuwijk J, Kant SG, et al. Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet. 2010;87:418–23.
Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. Eur J Hum Genet. 2012. doi:10.1038/ejhg.2011.258. ejhg2011258 [pii].
Goel R, Harsha HC, Pandey A, Prasad TS. Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. Mol Biosyst. 2012;8:453–63.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proc Natl Acad Sci U S A. 2007;104:8685–90.
Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score. Condel Am J Hum Genet. 2011;88:440–9.
Gotz A, Tyynismaa H, Euro L, Ellonen P, Hyotylainen T, Ojala T, Hamalainen RH, Tommiska J, Raivio T, Oresic M, et al. Exome sequencing identifies mitochondrial alanyl-tRNA synthetase mutations in infantile mitochondrial cardiomyopathy. Am J Hum Genet. 2011;88:635–42.
Grau D, Serbedzija G. Innovative strategies for drug repurposing. Drug Discovery Dev. 2005.
Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). Hum Mutat. 2000;15:57–61.
Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, de Vries P, de Reuver R, Wieskamp N, Mortier G, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet. 2010;42:483–5.
Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. Int J Med Inform. 2005;74:289–98.
Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A. 2010a;107:14621–6.
Iorio F, Isacchi A, di Bernardo D, Brunetti-Pierri N. Identification of small molecules enhancing autophagic function from drug network analysis. Autophagy. 2010b;6:1204–5.
Isidor B, Lindenbaum P, Pichon O, Bezieau S, Dina C, Jacquemont S, Martin-Coignard D, Thauvin-Robinet C, Le Merrer M, Mandel JL, et al. Truncating mutations in the last exon of NOTCH2 cause a rare skeletal disorder with osteoporosis. Nat Genet. 2011;43:306–8.
Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. Nature. 2001;409:853–5.
Johnson JO, Mandrioli J, Benatar M, Abramzon Y, Van Deerlin VM, Trojanowski JQ, Gibbs JR, Brunetti M, Gronka S, Wuu J, et al. Exome sequencing reveals VCP mutations as a cause of familial ALS. Neuron. 2010;68:857–64.
Junker BH, Koschutzki D, Schreiber F. Exploration of biological network centralities with CentiBiN. BMC Bioinf. 2006;7:219.
Kaimal V, Sardana D, Bardes EE, Gudivada RC, Chen J, Jegga AG. Integrative systems biology approaches to identify and prioritize disease and drug candidate genes. Methods Mol Biol. 2011;700:241–59.
Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006;34:D354–7.
Kann MG. Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform. 2007;8:333–46.
Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, et al. Predicting new molecular targets for known drugs. Nature. 2009;462:175–81.
King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science (New York, NY). 1975;188:107–16.
Kingsmore SF, Saunders CJ. Deep sequencing of patient genomes for disease diagnosis: when will it become routine? Sci Transl Med. 2011;3:87ps23.
Kinnings SL, Liu N, Buchmeier N, Tonge PJ, Xie L, Bourne PE. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. PLoS Comput Biol. 2009;5:e1000423.
Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.
Kohler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008;82:949–58.
Korstanje R, Paigen B. From QTL to gene: the harvest begins. Nat Genet. 2002;31:235–6.
Krawitz PM, Schweiger MR, Rodelsperger C, Marcelis C, Kolsch U, Meisel C, Stephani F, Kinoshita T, Murakami Y, Bauer S, et al. Identity-by-descent filtering of exome sequence data

identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet. 2010;42:827–9.
Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P. STITCH 3: zooming in on protein-chemical interactions. Nucleic Acids Res. 2012;40:D876–80.
Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.
Lalonde E, Albrecht S, Ha KC, Jacob K, Bolduc N, Polychronakos C, Dechelotte P, Majewski J, Jabado N. Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. Hum Mutat. 2010;31:918–23.
Lamb J. The connectivity map: a new tool for biomedical research. Nat Rev Cancer. 2007;7:54–60.
Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science (New York, NY). 2006;313:1929–35.
Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships. PLoS One. 2009;4:e4346.
Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25:2744–50.
Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res. 2012;40:e53.
Lin Z, Chen Q, Lee M, Cao X, Zhang J, Ma D, Chen L, Hu X, Wang H, Wang X, et al. Exome sequencing reveals mutations in TRPV3 as a cause of Olmsted syndrome. Am J Hum Genet. 2012;90:558–64.
Linghu B, Snitkin ES, Hu Z, Xia Y, Delisi C. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol. 2009;10:R91.
Lopez-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res. 2004;32:3108–14.
Mackay TF. Quantitative trait loci in Drosophila. Nat Rev. 2001;2:11–20.
Majewski J, Schwartzentruber JA, Caqueret A, Patry L, Marcadier J, Fryns JP, Boycott KM, Ste-Marie LG, McKiernan FE, Marik I, et al. Mutations in NOTCH2 in families with Hajdu-Cheney syndrome. Hum Mutat. 2011;32:1114–7.
Marchegiani S, Davis T, Tessadori F, van Haaften G, Brancati F, Hoischen A, Huang H, Valkanas E, Pusey B, Schanze D, et al. Recurrent mutations in the basic domain of TWIST2 cause Ablepharon Macrostomia and barber-say syndromes. Am J Hum Genet. 2015;97:99–110.
Masseroli M, Martucci D, Pinciroli F. GFINDer: genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. Nucleic Acids Res. 2004;32:W293–300.
Masseroli M, Galati O, Pinciroli F. GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. Nucleic Acids Res. 2005;33:W717–23.
Meester JA, Southgate L, Stittrich AB, Venselaar H, Beekmans SJ, den Hollander N, Bijlsma EK, Helderman-van den Enden A, Verheij JB, Glusman G et al. Heterozygous loss-of-function mutations in DLL4 cause Adams-Oliver syndrome. Am J Hum Genet. 2015;97:475–82.
Morrison AC, Voorman A, Johnson AD, Liu X, Yu J, Li A, Muzny D, Yu F, Rice K, Zhu C, et al. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet. 2013;45:899–901.
Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, Sougnez C, Garimella KV, Fisher S, Abreu J, Barry AJ, et al. Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N Engl J Med. 2010;363:2220–7.
Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010;42:790–3.
O'Connor KA, Roth BL. Finding new tricks for old drugs: an efficient route for public-sector drug discovery. Nat Rev Drug Discov. 2005;4:1005–14.

O'Sullivan J, Bitu CC, Daly SB, Urquhart JE, Barron MJ, Bhaskar SS, Martelli-Junior H, dos Santos Neto PE, Mansilla MA, Murray JC, et al. Whole-Exome sequencing identifies FAM20A mutations as a cause of amelogenesis imperfecta and gingival hyperplasia syndrome. Am J Hum Genet. 2011;88:616–20.

ODA. The Orphan Drug Act – implementation and impact. Department of Health and Human Services, Office of Inspector Journal. 2001.

Olatubosun A, Valiaho J, Harkonen J, Thusberg J, Vihinen M. PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat. 2012;33:1166–74.

Ortutay C, Vihinen M. Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. Nucleic Acids Res. 2009;37:622–8.

Padhy BM, Gupta YK. Drug repositioning: re-investigating existing drugs for new therapeutic indications. J Postgrad Med. 2011;57:153–60.

Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. Nat Genet. 2002;31:316–9.

Perez-Iratxeta C, Wjst M, Bork P, Andrade MA. G2D: a tool for mining genes associated with disease. BMC Genet. 2005;6:45.

Pierce SB, Walsh T, Chisholm KM, Lee MK, Thornton AM, Fiumara A, Opitz JM, Levy-Lahad E, Klevit RE, King MC. Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome. Am J Hum Genet. 2010;87:282–8.

Piro RM, Di Cunto F. Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J. 2012;279:678–96.

Puente XS, Quesada V, Osorio FG, Cabanillas R, Cadinanos J, Fraile JM, Ordonez GR, Puente DA, Gutierrez-Fernandez A, Fanjul-Fernandez M, et al. Exome sequencing and functional analysis identifies BANF1 mutation as the cause of a hereditary progeroid syndrome. Am J Hum Genet. 2011;88:650–6.

Pujol A, Mosca R, Farres J, Aloy P. Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci. 2010;31:115–23.

Rados C. Orphan products: hope for people with rare diseases. FDA Consum. 2003;37:10–5.

Ratbi I, Falkenberg KD, Sommen M, Al-Sheqaih N, Guaoua S, Vandeweyer G, Urquhart JE, Chandler KE, Williams SG, Roberts NA, et al. Heimler syndrome is caused by hypomorphic mutations in the peroxisome-biogenesis genes PEX1 and PEX6. Am J Hum Genet. 2015;97:535–45.

Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39:e118.

Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, Macii E, Benini L, Volinia S. TOM: a web-based integrated approach for identification of candidate disease genes. Nucleic Acids Res. 2006;34:W285–92.

Russ AP, Lampel S. The druggable genome: an update. Drug Discov Today. 2005;10:1607–10.

Sardana D, Zhu C, Zhang M, Gudivada RC, Yang L, Jegga AG. Drug repositioning for orphan diseases. Brief Bioinform. 2011;12:346–56.

Sasidharan Nair P, Vihinen M. VariBench: a benchmark database for variations. Hum Mutat. 2013;34:42–9.

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7:575–6.

Shihab HA, Gough J, Cooper DN, Day IN, Gaunt TR. Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics. 2013;29:1504–10.

Simpson MA, Irving MD, Asilmaz E, Gray MJ, Dafou D, Elmslie FV, Mansour S, Holder SE, Brain CE, Burton BK, et al. Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. Nat Genet. 2011;43:303–5.

Smith NG, Eyre-Walker A. Human disease genes: patterns and predictions. Gene. 2003;318:169–75.

Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Biol. 2010;6:e1000662.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13:2129–41.

Thornblad TA, Elliott KS, Jowett J, Visscher PM. Prioritization of positional candidate genes using multiple web-based software tools. Twin Res Hum Genet. 2007;10:861–70.

Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res. 2005;33:1544–52.

Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, Oti M, Lopez-Bigas N, Ouzounis C, Perez-Iratxeta C, Andrade-Navarro MA, et al. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Res. 2006;34:3067–81.

Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y. ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res. 2008;36:W377–84.

Turner FS, Clutterbuck DR, Semple CA. POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol. 2003;4:R75.

UniProt C. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014;42:D191–8.

UniProt C. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43:D204–12.

United States Food and Drug A. The Orphan Drug regulations. 1992.

van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. Eur J Hum Genet. 2003;11:57–63.

van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. Nucleic Acids Res. 2005;33:W758–61.

van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. Eur J Hum Genet. 2006;14:535–42.

Vissers LE, Lausch E, Unger S, Campos-Xavier AB, Gilissen C, Rossi A, Del Rosario M, Venselaar H, Knoll U, Nampoothiri S, et al. Chondrodysplasia and abnormal joint development associated with mutations in IMPAD1, encoding the Golgi-resident nucleotide phosphatase, gPAPP. Am J Hum Genet. 2011;88:608–15.

Wang JL, Yang X, Xia K, Hu ZM, Weng L, Jin X, Jiang H, Zhang P, Shen L, Guo JF, et al. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain. 2010;133:3510–8.

Wastfelt M, Fadeel B, Henter JI. A journey of hope: lessons learned from studies on rare diseases and orphan drugs. J Intern Med. 2006;260:1–10.

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34:D668–72.

Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36:D901–6.

Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. Mol Syst Biol. 2008;4:189.

Wu C, Gudivada RC, Aronow BJ, Jegga AG. Computational drug repositioning through heterogeneous network clustering. BMC Syst Biol. 2013;7 Suppl 5:S6.

Xu K, Cote TR. Database identifies FDA-approved drugs with potential to be repurposed for treatment of orphan diseases. Brief Bioinform. 2011;12:341–5.

Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics. 2006;22:2800–5.

Zhang M, Zhu C, Jacomy A, Lu LJ, Jegga AG. The orphan disease networks. Am J Hum Genet. 2011;88:755–66.

Zhu M, Zhao S. Candidate gene identification approach: progress and challenges. Int J Biol Sci. 2007;3:420–7.

Zhu C, Kushwaha A, Berman K, Jegga AG. A vertex similarity-based framework to discover and rank orphan disease-related genes. BMC Syst Biol. 2012a;6 Suppl 3:S8.

Zhu F, Shi Z, Qin C, Tao L, Liu X, Xu F, Zhang L, Song Y, Zhang J, Han B, et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. Nucleic Acids Res. 2012b;40:D1128–36.

Zuchner S, Dallman J, Wen R, Beecham G, Naj A, Farooq A, Kohli MA, Whitehead PL, Hulme W, Konidari I, et al. Whole-exome sequencing links a variant in DHDDS to retinitis pigmentosa. Am J Hum Genet. 2011;88:201–6.

Chapter 17
Toward Pediatric Precision Medicine: Examples of Genomics-Based Stratification Strategies

Jacek Biesiada, Senthilkumar Sadhasivam, Mojtaba Kohram, Michael Wagner, and Jaroslaw Meller

Abstract Next generation sequencing and low-cost genotyping technologies are opening up new avenues to translate genomics-based stratification and genetic variant information into improved care and personalized interventions in the clinic. Towards that goal, genome-wide variant information has become an important tool for cohort identification and stratification, phenotype-genotype association studies, discovery of disease markers, prediction of endo-phenotypes, and clinical decision

J. Biesiada, Ph.D. (✉)
Division of Biostatistics and Bioinformatics, Department of Environmental Health, University of Cincinnati College of Medicine,
Kettering Lab Building, 160 Panzeca Way, Cincinnati, OH 45267-0056, USA
e-mail: biesiajk@ucmail.uc.edu

S. Sadhasivam, M.D.
Department of Anesthesia, University of Cincinnati College of Medicine, Cincinnati Children's Hospital Medical Center,
3333 Burnet Avenue, ML-7024, Cincinnati, OH 45229-3039, USA
e-mail: Senthil.Sadhasivam@cchmc.org

M. Kohram, M.S.
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center,
3333 Burnet Avenue, ML-7024, Cincinnati, OH 45229-3039, USA
e-mail: Mojtaba.Kohram@cchmc.org

M. Wagner, Ph.D.
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center,
3333 Burnet Avenue, ML-7024, Cincinnati, OH 45229-3039, USA

Department of Pediatrics, University of Cincinnati College of Medicine,
Cincinnati, OH, USA
e-mail: Michael.Wagner@cchmc.org

J. Meller, Ph.D.
Departments of Environmental Health, Pediatrics and Biomedical Informatics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center,
3333 Burnet Avenue, ML-7024, Cincinnati, OH 45229-3039, USA
e-mail: jarek.meller@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_17

support. This chapter focuses on the use of genetic variant information in the context of pediatric autoimmune diseases and pain management in a pediatric surgery setting. Genome-wide variant detection and discovery, as well as targeted gene sequencing approaches are discussed through the lenses of the resulting informatics challenges, implied tailored research informatics solutions, and integration with clinical informatics systems. These challenges and solutions are illustrated using three specific applications, namely: (i) cohort stratification analysis; (ii) prediction of classical HLA alleles from variant data in the context of pediatric autoimmune diseases; and (iii) predictive decision models for the management of surgical pain and opioid-related adverse outcomes in children.

Keywords Alleles • Decision support • HLA • Molecular modeling • Pain • Opioids • Genetic variants

17.1 Introduction

The rapid adoption of next-generation DNA sequencing and its integration with clinical informatics systems and analytical systems are increasingly enabling personalized clinical interventions (Holmes et al. 2009; Chan and Ginsburg 2011). Research and clinical centers, including those focused on pediatric health, have been investing heavily in the development of necessary informatics infrastructure, including scalable databases that combine different types of information, analytical and data mining capabilities, as well as better interactive platforms for cross-disciplinary basic and translational research.. This chapter focuses on genetic stratification strategies using variant information obtained by next generation sequencing, genome-wide genotyping arrays as well as re-sequencing arrays (LaFramboise 2009; Schaaf et al. 2011). These platforms readily provide information about genetic variation that can subsequently be combined with demographic, phenotypic and other data to facilitate both basic and translational clinical research (Chou et al. 1987; 2006a, b; Diatchenko et al. 2005; Wellcome Trust Case Control Consortium 2007; de Bakker et al. 2006). Informatics challenges and solutions pertaining to the use of genetic variant information are discussed, with several case studies used to illustrate applications in basic research and clinical settings.

We start by describing the development and use of integrated databases, computational analysis pipelines and platforms for joint analysis of variant information and clinical data. These databases are in use at our institution and enable seamless integration with clinical informatics systems. Emphasis is placed on informatics challenges related to: the evolving nature and increasing complexity of the data, as new re-sequencing platforms are being adopted; data are generated by multiple centers and must be combined to increase the statistical power of analyses; and ongoing evolution of sequencing platforms and processing pipelines that may present data in new formats or with different attributes.

The first application (Sect. 17.3) concerns the prediction of Human Leukocyte Antigen (HLA) alleles (Lechler and Warrens 2000; Marsh et al. 2010a, b; Aureli et al. 2008; Orozco et al. 2008; Sampaio-Barros et al. 2008; Robinson et al. 2003, 2009; Leslie et al. 2008) from SNP variant data. Reliable HLA allele prediction (or imputation) methods have recently been developed (Leslie et al. 2008; Dilthey et al. 2011), offering significant advantages in on-going studies of pediatric autoimmune disorders. At the same time, it adds to the complexity of informatics solutions that are required to fully capitalize on these developments. A number of additional layers, including carefully assessed and unified quality control protocols, population stratification and haplotype reconstruction (de-phasing) approaches, and multiple filters and predictors must be integrated to achieve robust accuracy and high confidence in predictions. Possible applications to donor matching with further progress in this field are also discussed.

The second application (Sect. 17.4) deals with decision support systems that are being developed to minimize opioid-related adverse effects, while managing perioperative pain in children. Children are at higher risk of inadequate pain relief and serious side effects from morphine, for example in the context of pediatric surgery pain management (Caldas et al. 2004). The ability to personalize the use of morphine in order to maximize pain relief while minimizing its adverse effects has clear implications for public health. Efforts to develop decision support systems for such personalized intervention provide a case study for applications of informatics in translational research.. Here, we focus on challenges related to developing robust and accurate decision rules, predictive genetic signatures and risk-assessment tools that combine genotypes and clinical data in the context of population stratification and gene-gene interactions.

17.2 Integrating Large Variant Data with Phenotypic Information

Advances in commercially available sequencing technology (in particular by Illumina and other next-generation sequencing outfits) have facilitated the use of whole-exome (WES) and whole-genome sequencing (WGS) to investigate links between specific genetic variants with human disease. Prior to the maturing of sequencing technology, whole-genome genotyping platforms were already able to probe individuals' genomes for single nucleotide variants at millions of loci for use in genome-wide association studies (GWAS). Both of these types of studies use allele frequencies in populations under considerations to find correlations between a phenotype (disease status) and individual genomic variants (typically single 87 nucleotide polymorphisms or SNPs), but also copy number variants (CNVs) and insertion-deletions (indels) (Fig. 17.1).

The large scale of the source data (a typical raw fastq data file for WGS can have a size of tens of gigabytes) and, especially in the case of GWAS, the need for data

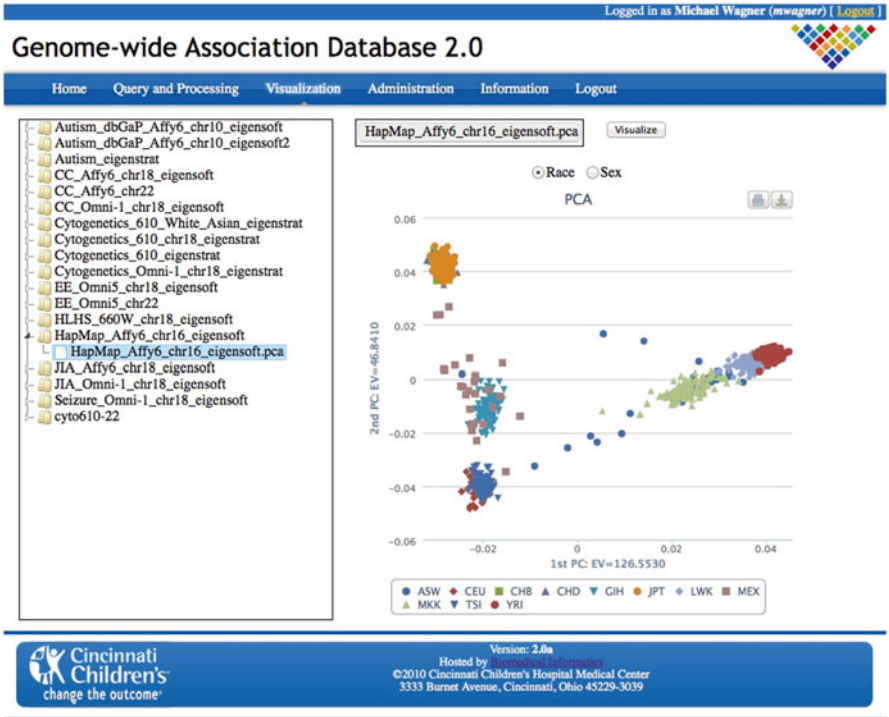


Fig. 17.1 Visualization of population sub-structure using Eigenstrat on HapMap3 data. All SNPs on chromosome 16 were used on this Affymetrix 6 data set. HapMap3 populations (including European ancestry CEU and TSI cohorts) are shown using two main PCA components. In-browser visualization is facilitated by a library from HighCharts.com (The International HapMap Consortium 2003)

on very large cohorts (often tens of thousands of individuals), require efficient data management and pose challenges in data analyses. While a number of very efficient and high-quality analysis software packages are available for both sequence processing (e.g., BWA (Li and Durbin 2009) for alignment, GATK (McKenna et al. 2010) for variant calling, etc) and GWAS analysis (e.g., PLINK, SNPTEST, LAMP (Purcell et al. 2007; Marchini and Howie 2010; Li et al. 2005)), the challenge of transparently managing genome-wide variant information in conjunction with demographic and phenotypic variables is less easily solved with off-the-shelf tools.

17.2.1 Storing and Accessing Raw Sequencing and Genotyping Data

Both raw sequencing and genotyping data typically must undergo iterative rounds of filtering and quality control, which usually results in many versions of the same data, which in turn create challenges in documentation and tracking data provenance. This has motivated us to improve processes for managing large

genome-wide data sets, both in cases of raw sequencing data and raw genotyping data. For the latter, and in order to provide a central, web-accessible repository for, e.g., CEL files for Affymetrix and idat files for Illumina platforms, we developed an application (accessible to registered users at http://research.cchmc.org/genotyping), which allows investigators to perform the following functions:

- Upload new data files that conform to a certain file format. All files are parsed, any metadata contained in the file is parsed out by the system.
- Control access to data, based on permissions granted by the owner of the data
- Tag data files with keywords and controlled vocabulary
- Upload text files or spreadsheets with supplemental information about the primary data
- Use keyword, controlled vocabulary and/or free-text searches to retrieve and download primary data
- Use the interface to launch processing (genotype calling) algorithms on a Linux cluster computer, thus enabling access to very high-performance computing power for basic researchers otherwise unfamiliar with computing tools or languages. In particular, we have implemented the CRLMM (Ritchie et al. 2009) approach to genotype calling, and we make the Affy Power Tools available to our users as well.

A similar portal also exists for fastq files which contain raw sequence reads for either Whole Exome or Whole Genome Sequencing data, although this is being subsumed by a more sophisticated infrastructure that is being developed at CCHMC under the auspices of the Center for Pediatric Genomics (a.k.a. CpG). Our VIVA portal, which is under active development as of the writing of this chapter, will provide a web-based interface for investigators to upload saw sequencing data. All data from the CCHMC Sequencing Core will be uploaded seamlessly and is analyzed using a standard, state-of-the-art computational pipeline build around the Genomic Analysis Tool Kit (GATK, (McKenna et al. 2010)). VIVA will hold both raw (fastq), aligned (bam) and called variant (vcf) files and will have many of the same features described above for the case of genotyping data. In particular because of the large data size we feel strongly that our approach of providing user-friendly, centralized data stores for genomic data is essential for efficient research data management.

17.2.2 Integrating Genotype and Phenotype Data Flows

Unlike the large raw data files, genotype call files (or, in the case of sequencing data, vcf files) are much smaller in size. These data types provide the basis used for statistical analysis and interpretation in the context of phenotypic information. To facilitate analysis of next generation sequencing data at CCHMC, we recently developed a bioinformatics NGS pipeline called the Cincinnati Analytical Suite for Sequencing Informatics (CASSI, http://cassi.research.cchmc.org) to integrate the data storage, filtering, and annotation through a web-based interface (Patel et al.

2014). CASSI analysis pipelines are run on the CCHMC 1300+ core Linux-based computational cluster. It leverages existing open source VCF file parsers and annotation tools including VCF tools (Danecek et al. 2011), ANNOVAR (Wang et al. 2010), the UCSC Genome Browser (Kent et al. 2002), Exome Variant Server, and NIH-hosted genomic data repository dbGAP. After using standard, GATK-based algorithms for the initial read processing and alignment, three quality control measurements are used: (1) read depth (number of sequencing reads that contain the variant); (2) genotype quality score (GQ) of each of the variants called and (3) the expected alternate allele ratio (alt ratio) for a particular genotype. The alternate allele ratio is the proportion of the number of reads with the alternate allele at a position relative to the total number of reads at that same position. Our previous work showed that filtering the variants on these measurements removed ~95 % of the sequencing artifacts and Mendelian errors in trios while retaining 80 % of the called variants (Patel et al. 2014). After filtering for high quality variants, the samples will be then scanned for amino acid altering variants (non-synonymous, splicing, insertions, deletions, and variations that alter initiation codons or stop codons) using the UCSC genome browser build 37 human Reference Sequence Gene table. Rare and novel variants will be identified by filtering against the 1000 (The 1000 Genomes Project Consortium 2015), the NHLBI Exome Sequence Project and an internal allele frequency table of 312 whole exomes analyzed at CCHMC as it is assumed that common variants represent harmless variations. Variant *annotation* is carried out by state-of-the-art tools such as ANNOVAR (Wang et al. 2010) and the Variant Tools package (Danecek et al. 2011), which offer methods of examining their functional consequence on genes (e.g., missense, nonsense, splicing disruption, frame shifting indel, and non-frame shifting indel). Variants are annotated with chromosome, position, minor allele frequency, gene name (www.Genecards.org), transcript, and protein ID and amino-acid position. A schematic outline of the overall strategy and experimental procedures proposed to achieve this goal is shown in Fig. 17.2. All steps of the processing pipeline are being implemented using the LONI pipeline software (Dinov et al. 2009). CASSI provides a web-accessible, user-friendly and secure interface that lets investigators manage all sequencing data along with clinical information.. It interfaces directly with the LONI pipeline engine via a webstart module: search results returned by the web interface are thus seamlessly staged for processing with the bioinformatics tools mentioned above on the Linux cluster. Previously published work (Patel et al. 2014) indicates that our CASSI pipelines are robust with regard to sequencing accuracy and identification of candidate variants. Individual and summary reports will be generated for all candidate variants.

For GWAS data, we had previously developed an application application, which we call gwadb (Genome-Wide Association Database), and which consists of a user-friendly, web-based interface to a relational database holding all relevant data (in particular, genotype calls) and metadata for GWAS analyses. Our aim was to enable clinical and translational researchers to perform GWAS analyses and related tasks on well-curated, well-documented data without requiring knowledge of Linux command lines. We accomplish this by storing raw or minimally processed genotyping

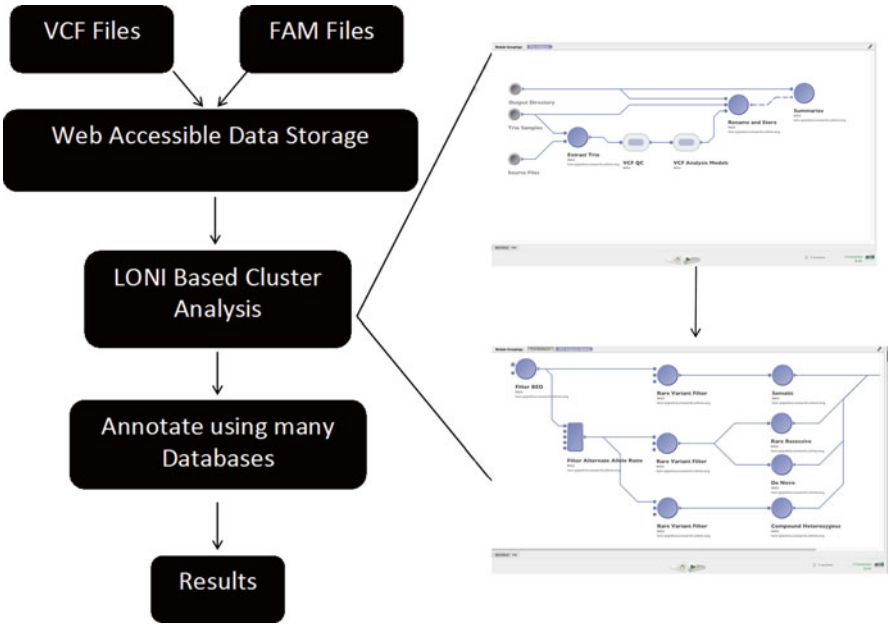


Fig. 17.2 Graphical representation of the Cincinnati Analytical Suite for Sequencing Informatics. Sequence data are parsed and stored in web accessible data storage. The LONI based analysis allows users to analyze data through the established pipeline for identifying relevant variants. These pipelines can also be changed to accommodate specific analyses. After variants are annotated using many databases, the results are versioned, saved in the database, and available for downloading

data in the same relational database as demographic and phenotypic information. We enable our users not only to download any data in a number of commonly used file formats, but also to process these data through a number of pre-defined, parameterized workflows which run on a Linux cluster. The capability of executing these complex and often CPU-time and memory-intensive processes as a batch job on a cluster is essential for GWAS-type analyses, as the data sizes often make local processing close to impossible.

Currently the processing workflows that are directly callable from gwadb's web interface include commonly used tools such as PLINK (Purcell et al. 2007) (for quality control, case-control analyses, basic data management etc.), Eigenstrat (for cohort variability visualization), IMPUTE2 (Howie et al. 2009; Marchini et al. 2007) (for imputation of genotypes to the 1000 Genomes reference set), KING (Manichaikul et al. 2010) (for automated pedigree imputation based on genotypes). All of these are essential to perform GWAS studies. PLINK is widely used in the community and provides a large number of functions to summarize, manage, filter and analyze large-scale SNP data, it has de facto become a standard tool in any analysis. Association p-values can be uploaded directly into a track of a local instance of the UCSC Genome Browser (Kent et al. 2002) and further processed and visualized.

The analysis of ancestry with tools such as Eigenstrat (Price et al. 2006) is equally essential, not only for cohort stratification, but also as a quality control tool since it in effect allows for verification of self-reported ancestries. Imputation is a crucial tool when performing analyses across different genotyping platforms, and because of its computational cost, it is especially convenient for our end-users to run IMPUTE2 (Howie et al. 2009; Marchini et al. 2007) on our computational cluster from gwadb. Finally, GWAS studies generally assume that all data used come from unrelated subjects, and KING (Manichaikul et al. 2010) is a convenient, fast computational tool that can verify kinships and uncover previously unknown familial relationships. This is especially useful when merging different cohorts in order to verify that no one is enrolled in both cohorts at the same time. As new applications and extensions are considered, more software tools can easily be incorporated into gwadb due its flexible design, modular architecture, and the integration with a Linux computational cluster, which enables parallel large scale computation in the batch mode, as well as on-the-fly analysis whenever applicable.

Gwadb is implemented using open-source tools such as php/mysql. Furthermore all communication to the server is encrypted with SSL and strict guidelines are enforced concerning user names and passwords, thus ensuring security and compliance with security standards. These portals represent complete suites that integrate the management of disk-based storage for primary data, the management and integration of metadata about the primary data with clinical data points, as well as the processing of these data using state-of-the-art algorithms and computational equipment in a user-friendly web interface.

17.3 Prediction of HLA Alleles from SNP Data

Human Leukocyte Antigen (HLA) locus on chromosome 6 encodes a number of highly polymorphic genes that play an essential role in immune system responses (de Bakker et al. 2006; Lechler and Warrens 2000; Marsh et al. 2010a, b; Aureli et al. 2008; Orozco et al. 2008; Sampaio-Barros et al. 2008; Robinson et al 2003, 2009; Leslie et al. 2008). In particular, Major Histocompatibility Complex (MHC) class I and class II genes, including HLA-A,-B, -C and DRB1, DQB1, DPB1, respectively, are well known components of the immune system. The function and medical relevance of these genes stem from their ability to bind and selectively recognize peptides derived from pathogens as well as self-peptides. The latter aspect of their function contributes to the role of MHC region as an important susceptibility locus in autoimmune diseases (de Bakker et al. 2006; Lechler and Warrens 2000). In particular, multiple associations between HLA alleles and childhood autoimmune disorders have been observed, including type 1 diabetes, celiac disease and Juvenile Idiopathic Arthritis (JIA). For example, HLA-B gene variant (allele) HLA-B35 and HLA-DRB1 gene variants (alleles) DRB1*1101/1104 have been associated with susceptibility to some JIA subtypes, whereas HLA-DR4 appears to be a protective allele. Consequently, HLA loci and their haplotypes are an important

component of disease association studies in the context of JIA and other autoimmune diseases (Aureli et al. 2008; Orozco et al. 2008; Sampaio-Barros et al. 2008). Recently, the 1000 Genome Project and related large scale sequencing efforts provided further evidence of high degree of genetic variability within the HLA locus, with many new alleles, ancestry-specific linkage disequilibrium (LD) structures, and association patterns being mapped, as a result of these efforts.

HLA alleles can be typed by direct DNA sequencing of variable regions of HLA genes, such as exon 2 of DRB1, which encode parts of the protein involved in peptide recognition. This approach is non-trivial due to highly variable and poly-genic nature of the HLA locus, often requiring disambiguation and implying the use of specialized sequencing centers. As an alternative, one can take advantage of single nucleotide polymorphism (SNP) genotype data that can be easily (and inexpensively) generated using SNP microarrays. These "SNP chips" can be used to resequence single letter variants in the whole genome, or they can be tailored to include a specific panel of SNPs, e.g., sampling the HLA locus (LaFramboise 2009; Schaaf et al. 2011). In particular, the existence of strong linkage disequilibrium between MHC SNPs and HLA alleles suggested that SNP tagging (or SNP-based imputing) could be an attractive (and inexpensive) alternative to traditional HLA typing (de Bakker et al. 2006; Leslie et al. 2008; Dilthey et al. 2011).

Efficient methods for SNP genotyping have significantly impacted studies on autoimmunity. Large-scale genotyping projects using SNP arrays are providing rich data on variation within the extended MHC region in different populations, adding to 1000 Genome Project and other efforts to map human genetic variation. SNP arrays can also be used to detect copy number variants (CNVs), which provides additional advantages (CNV-based stratification), especially for highly poly-genic regions, such as HLA.

17.3.1 Linkage Disequilibrium

Extensive LD patterns present within the MHC region are illustrated in Fig. 17.3, using data on a cohort of about 500 healthy controls of Caucasian descent from the Midwestern United States. This and similar data sets are made available to researchers through the gwadb database described in Sect. 17.2. SNP genotypes in the region were obtained using the Affymetrix 6 platform, whereas high resolution (4 digit) HLA genotypes were obtained by direct DNA sequencing of the variable exons, following standard protocols. PHASE version 2.1 (Marchini et al 2006; Stephens and Donnelly 2003; Stephens et al. 2001) was used to reconstruct the haplotypes, which was facilitated by the use of gwadb and the Linux computational cluster. The strength of correlation between individual (biallelic) SNPs and (multiallelic) HLA was then quantified using the relative information, following deBakker (de Bakker et al. 2006). Different colors represent the extent of LD for each of 8 HLA genes considered here (Fig. 17.4).

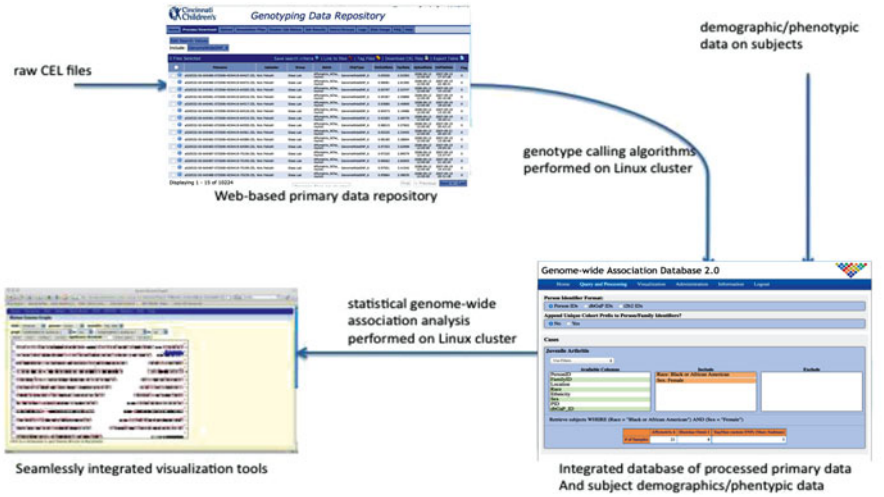


Fig. 17.3 Overview of the genotype/phenotype data flow through web-based repositories, including the raw data repository (*top left*) and gwadb (*bottom right*). We emphasize the ability to spawn computational analysis jobs (e.g., for genotype calling and GWAS analysis) on a high-performance Linux cluster

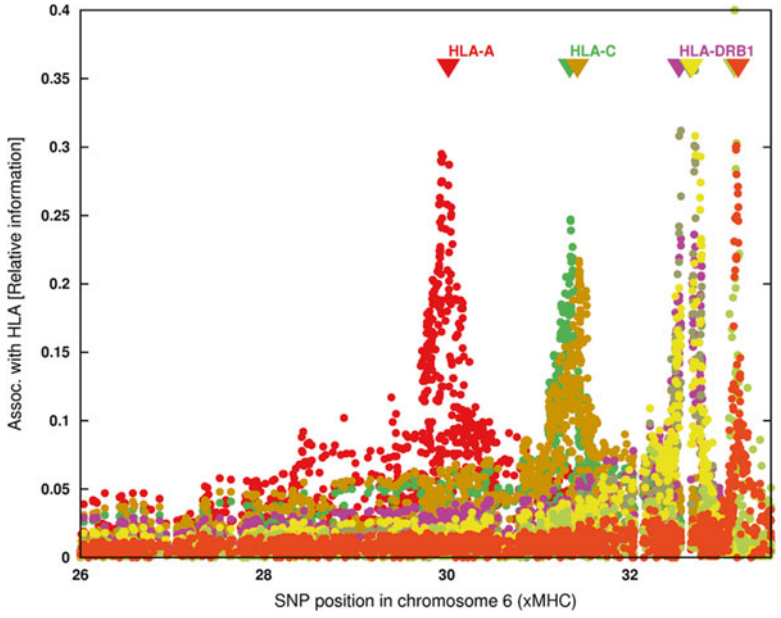


Fig. 17.4 Extensive linkage disequilibrium within the MHC region

As can be seen from the figure, there is a broad peak of LD around each gene. Consequently, SNPs located near (and in majority of cases outside of) the respective HLA genes can be used to impute their alleles based on these strong LD patterns. Such predictions for cohorts of Caucasian ancestry have been demonstrated to reach greater than 90 % accuracy at the level of 4 digit alleles for all except one gene considered here, the exception being DRB1 for which greater than 85 % accuracy has been reported (Leslie et al. 2008; Dilthey et al. 2011). Prediction accuracy for other ancestries, including some rare populations studied as part of the 1000 Genome Project, have recently been assessed systematically as well, suggesting similar trends and pointing out the need for representative training cohorts and careful validation (Pappas et al. 2016). It should be noted that HLA alleles are typically typed with 4-digit resolution for association studies and transplant donor matching.

17.3.2 Prediction of HLA Alleles

Reliable prediction of HLA alleles from SNP data can be obtained by exploiting the presence of linkage disequilibrium (non-random associations) between SNPs within MHC and particular HLA alleles, and by applying advanced statistical and machine learning methods (see Fig. 17.5). Integrating such predictions with platforms for SNP data analyses, such as gwadb, can greatly streamline GWAS studies, especially those on autoimmune disorders. Using this approach, analysis of associations with HLA alleles can be performed by applying SNP-based imputation, without additional sequencing of HLA genes. A number of studies successfully utilizing this approach have been published in recent years. Applications of SNP-based imputation of HLA alleles should take into account:

- Differences between SNP arrays, including general genome-wide (e.g., Affymetrix 6.0) or tailored arrays (e.g., Immunochip), which may require between platform imputation and/or careful selection of SNP subsets to be used for the prediction
- Limitations in accuracy of current methods that at present cannot fully replace HLA typing (especially for cases predicted with lower confidence)
- Careful assessment and proper use of population stratification, while excluding individuals of mixed ancestry.

With regards to population stratification, as demonstrated by recent benchmarking efforts and limitations of ancestry-independent universal predictors (Pappas et al. 2016), extensions to other than White/CEU populations may require extensive data on specific cohorts and (re-) training of HLA allele predictors. This observation can be extended to relatively rare (overall) classes that nevertheless can play in important role in some specific autoimmune disorder (e.g., as a high risk allele). Representative cohorts for that particular disease (and risk allele) may be required to enable robust extrapolation from training examples and sufficient accuracy for the allele of interest.

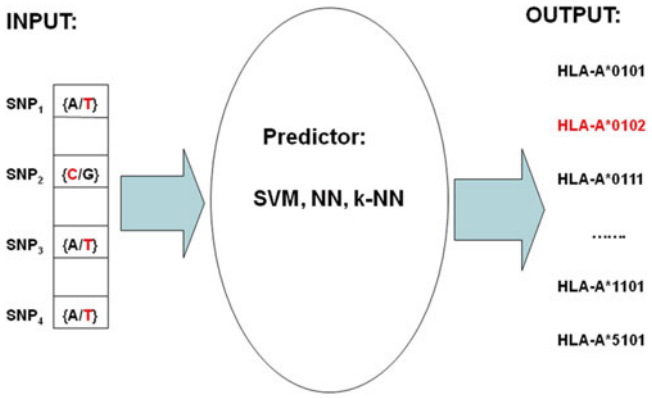


Fig. 17.5 HLA allele prediction from SNP data can be improved by using statistical and machine learning methods, such as Support Vector Machines (SVM) or neural network (NN); it requires carefully designed, properly stratified and sufficiently large high quality training and validation sets

Prediction of HLA alleles from SNP data starts from haplotype reconstruction, which allows one to assign individual HLA alleles to chromosomes reconstructed from SNP genotypes (see Fig. 17.5). Two different options are available for that purpose within our platform, integrating two widely used and well benchmarked packages for population-based haplotype reconstruction: Impute2 (Howie et al. 2009; Marchini et al. 2007) and Phase (version 2.1) (Stephens and Donnelly 2003; Stephens et al. 2001; Li and Stephens 2003). In case of the latter, which is in general more accurate, but also computationally more demanding, reconstruction can be performed within blocks of up to 400 SNPs (depending on the size of the cohort), spawning multiple jobs on the cluster for multiple blocks. Tagging SNPs (Howie et al. 2009; Marchini et al. 2006, 2007) across the whole region can be included to enable subsequent "stitching" to obtain partial haplotypes.

In summary, integrated informatics solutions for storing, processing and analyzing SNP and other sequencing data from multiple clinical studies can greatly facilitate efforts to extend and improve HLA allele prediction. Such databases and tools provide consistent quality control criteria and protocols, as well as access to data potentially pertaining to cohorts of different ancestries that can be used to assess, refine and re-train current methods. In particular, further improvements in terms of the accuracy can be obtained by combining publicly available data with local cohorts being collected and studied at multiple clinical centers. These efforts will further contribute to association studies by providing HLA allele assignment directly from SNP chip data, with impact on diagnosis, outcomes and personalized therapeutic interventions.

17.4 Variant-Based Models for Clinical Decision Rules

This part of the chapter focuses on efforts to improve quality of care in the context of pediatric surgery. One of the goals is to develop accurate models for the prediction of clinical responses to opioids, using genetic variant information in conjunction with clinical and demographic data. Safe and effective analgesia is an important medical and economic problem (Caldas et al. 2004; Duedahl and Hansen 2007). A significant fraction of approximately five million children, who undergo a painful surgery in the US each year, experience inadequate pain relief and serious opioid-related side-effects (Cepeda et al. 2003; Sadhasivam et al 2012; Esclamado et al 1989).

The efforts to personalize pain management in children build on the growing understanding of how genes and their specific variants influence postoperative pain and the occurrence of serious side effects, such as respiratory depression. Multiple genes involved in opioid metabolism, receptor signaling and transport have been associated with inter-individual variability in clinical response to morphine in small-scale studies of adults (Chou et al. 2006a, b; Diatchenko et al. 2006, 2007; Klepstad et al. 2004; Nackley and Diatchenko 2010; Nackley et al. 2007; Oertel et al. 2006; Rakvåg et al. 2005, 2008, Coller et al, 2006). Much of this variability in pain perception can be explained by SNPs in catechol-O-methyltransferase (COMT), which is a key regulator of pain perception. Genetic variants of mu opioid receptor (OPRM1), the main target for many opioids including morphine, have also been associated with increased pain sensitivity. Adult carriers of the GG variant of OPRM1 SNP A119G require a 2–4 times higher dose of morphine than AA variants. In addition, specific variants of the ATP Binding Cassette B1 (ABCB1) gene that encodes a transporter protein have been shown to modulate cerebral pharmacokinetics of morphine, thereby to change its analgesic and side effects (Chou et al. 2006a, b; Diatchenko et al. 2006, 2007; Klepstad et al. 2004; Nackley and Diatchenko 2010; Nackley et al. 2007; Oertel et al. 2006; Rakvåg et al. 2005, 2008, Coller et al. 2006).

With the goal of improving and personalizing the postoperative care and pain management in children, we have initiated a systematic effort to identify genetic variants underlying clinical responses to opioids. Our initial cohort included 199 children undergoing tonsillectomy, of which 39 were African-American and 99 were classified as suffering from obstructive sleep apnea (Esclamado et al. 1989). A standardized protocol allowed for unambiguous assignment and accurate quantification of the observed phenotypes. Genotype data were collected using a specialized SNP panel. The candidate genes included ABCB1, COMT, OPRM1, FAAH, ADRB2 and a number of other genes that were chosen based on their allele frequencies and clinical evidences of important associations in adults with opioid analgesic and adverse effects (Chou et al. 2006a, b; Diatchenko et al. 2006, 2007; Klepstad et al. 2004; Nackley and Diatchenko 2010; Nackley et al. 2007; Oertel et al. 2006; Rakvåg et al. 2005, 2008, Coller et al. 2006).

Preliminary statistical analyses revealed several significant associations between genetic and non-genetic factors and postoperative opioid adverse effects and inadequate analgesia. In particular, TT genotype of the ABCB1 SNP rs1045642 (C3435T) was found (after adjusting for obstructive sleep apnea) to be associated with a higher risk of morphine induced respiratory depression than CC genotype. Specifically, in ABCB1 TT genotype, resting minute ventilation (MV) after morphine decreased by 47.5 % compared to only 19.4 % in CC and CT genotypes ($p < 0.05$). A number of other, relatively weak associations were found with SNPs in FAAH, COMT and other genes, as well as indications of epistatic interactions between ABCB1 and FAAH and ABCB1 and ADRB2. We have performed a systematic multivariate analysis of associations between gene-gene interactions and other confounding factors (such as race, age, body mass index, etc.) and the respective outcomes, using standard CART and C4.5 decision trees [58–59]. We have also used other data mining and machine learning approaches to provide further insights into the problem (and improve prediction accuracies), including unsupervised clustering analyses to identify genetic "risk signatures" of opiod adverse effects.

It is readily apparent that informatics challenges and solutions for this application share commonalities with those discussed in the context of HLA allele prediction from SNP data. At the same time, this application illustrates other issues related to the use of clinical decision support systems. See Chap. 9 for a more extensive discussion of clinical decision support, an active area of translational research in pediatrics.

17.4.1 Decision Trees

Decision trees are a standard machine learning technique for multivariate data analysis and classification (Hastie et al. 2009; Witten and Frank 2005; Hothorn 2010). Decision trees can be viewed as a recursive partitioning approach, in which data is hierarchically divided into strata by simple logical rules. The advantage of decision trees is their simplicity, ability to handle both categorical and numerical variables as well as missing values, robustness to outliers and scaling, and the ability to combine feature selection with stratification and classification. Decision trees can also be used to derive easy to interpret and intuitive rules for decision support systems. Here, we use decision trees to select and combine the most predictive SNPs with demographic, clinical and other input features into simple logical rules that enable robust and accurate point-of-care prediction of inadequate pain relief and opioid-related adverse effects.

Different strata with distinct patterns of such interactions (with ABCB1 playing prominent role in some strata) were identified. Our preliminary data, which will guide the design of larger studies in the future as is typical in translational research, suggest that African American children had inadequate pain control and Caucasian children had a higher incidence of adverse effects from similar doses of morphine (Sadhasivam et al. 2012). Concordant differences in allelic frequency of ABCB1

(and other genes) were observed. For example, TT genotype of ABCB1 SNP rs1045642 that predisposes children to opioid induced respiratory depression (>4 fold higher incidence than CC genotype) was found with 27 % frequency in Caucasian children, as compared to 2–3 % in African-American children.

Here, we briefly describe the results obtained using standard CART and C4.5 decision trees, as implemented in R and Weka, respectively (Therneu and Atkinson 2009; Quinlan 1986; Biesiada et al. 2014). These methods were first used to identify and analyze potential patterns of gene-gene interactions and other factors predictive of pain sensitivity and inadequate pain relieve. The increased pain sensitivity phenotype is defined here by the need to administer a post-operative analgesic (PA). By contrast, the low pain sensitivity class is defined by no need for post-operative intervention in the form of additional analgesic. This analysis is summarized as an "efficacy" tree shown in Fig. 17.6, panel A, for the post-operative analgesic use vs. no intervention (noi) classification problem, using SNP data and covariates such as race and obstructive sleep apnea (OSA). Leaves (nodes) of the tree that represent strata with increased risk of inadequate pain relieve are highlighted in red, whereas those with relatively lower risk in green. The number of patients in each of the two classes is shown in each node of the tree to illustrate the results of recursive partitioning of the overall data set into subsequent subsets (strata).

Consistent with known associations between African-Americans, OSA and higher pain sensitivity, race is found to be the most discriminating feature, providing the first split of the overall cohort of 199 patients into two branches. Among Caucasian children, further strata are defined by specific polymorphisms in GCH1 (which was implicated as potential modifier of pain sensitivity and persistence) and interactions involving ABCD1/MC1R, ADRB2, and DRD2, or FAAH, DRD2, ABCD1/MC1R and TRPA1, in the two main sub-branches shown in Fig. 17.7 as

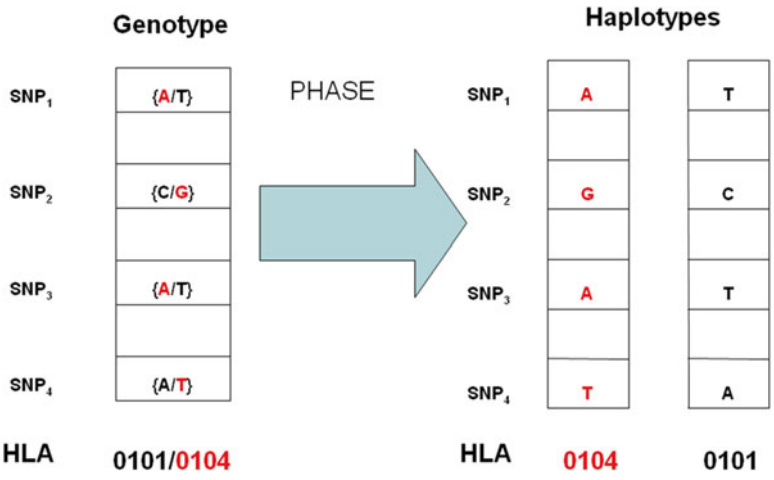


Fig. 17.6 Haplotype reconstruction for HLA prediction

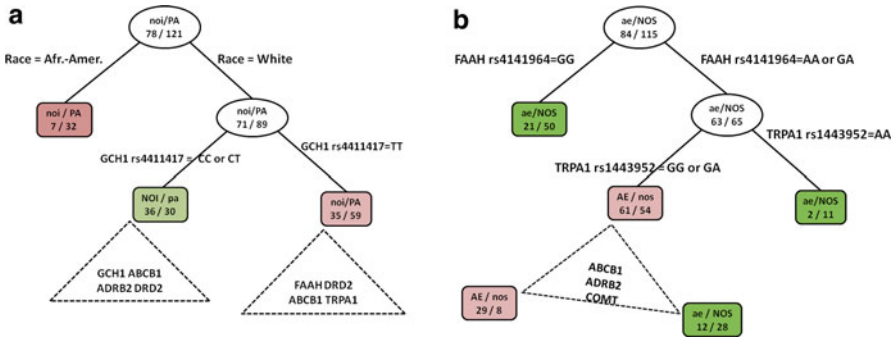


Fig. 17.7 Decision tree based prediction models for inadequate pain relief (efficacy tree, panel A) and adverse effects (safety tree, panel B); simple decision rules to be applied in clinical settings follow. (**a**) Efficacy Tree: no intervenation (noi) vs. the need for an additional post-operative analgesic (pa). (**b**) Safety Tree: adverse effects (ae) vs. no adverse effects, referred to as no symptoms (nos)

dotted triangles. The top part of the "efficacy" decision tree included in Fig. 17.7 panel A can be interpreted in terms of logical rules as follows:

IF (Race = African-American)

THEN ***high risk*** *of increased pain sensitivity and inadequate pain relieve*;

IF ((Race = White) AND (GCH1 rs441417 = TT))

THEN ***moderate risk*** *of increased pain sensitivity and inadequate pain relieve*;

IF ((Race = White) AND (GCH1 rs441417 = CC OR CT))

THEN ***low risk*** *of increased pain sensitivity and inadequate pain relieve.*

The structure of the tree and the resulting rules that could be used for point of care decision support systems capture a relatively strong correlation between race and increased pain sensitivity (82 % of African-American children included in the cohort required additional post-operative analgesic). At the same time, Caucasians could be further stratified by GCH1 and other genes, indicating epistatic effects. The C allele of GCH1 is associated with somewhat decreased pain sensitivity: about 63 % of Caucasian children with homozygous TT genotype required additional post-operative analgesic.

Additional analysis of factors predictive of adverse effects are summarized in the form of a "safety" tree (Fig. 17.7 panel B) for the classification of adverse effects, including respiratory depression, nausea and vomiting, and over-sedation (referred jointly to as AE) vs. the other (no adverse effects) class. Leaves of the tree that correspond to strata with increased risk of AE are highlighted in red. Only two main leaves that can be classified more easily by combinations of ABCB1, ADRB2, COMT and FAAH polymorphisms are shown explicitly within the middle branch.

Consistent with previous univariate analysis, GG genotype at FAAH rs4141964 SNP was found to be protective, whereas the other two genotypes in interaction with TRPA1 rs1443952 genotypes GG or GA (and specific further interactions with ABCB1, ADRB2, COMT and FAAH) carry increased risk of adverse effects.

Obstructive sleep apnea is over-represented among African-American patients, who in turn are more likely to be at a higher risk of inadequate pain relief, as also indicated by "efficacy" tree. OSA and race represent important covariates to be further studied using larger cohorts. This is further highlighted by the finding that ADRB2 and FAAH gene polymorphisms can predict race with approximately 80 % accuracy. However, OSA itself seems to have a relatively strong, although poorly understood, genetic component. We found that OSA can be predicted using SNPs in just two genes (ADRB2 and ABCB1) with greater than 70 % accuracy, compared to 50 % for a baseline classifier. It should be emphasized that further analysis of classification trees, stability of feature selection and the observed strata, as well as the accuracy of the resulting decision rules must be carefully assessed by study of larger numbers of patients from diverse populations. Preliminary data, such as presented here, guide the design of larger, more definitive, and more costly studies, as is typical in translational research. As with other machine learning methods, a common strategy to assess the overall accuracy and stability of decision trees, and to control the risk of overfitting is to use a comprehensive cross-validation approach, in which data is randomly split into training and validation subsets (Hastie et al. 2009; Witten and Frank 2005; Hothorn 2010). This approach works well with sufficiently large sample sizes so that different strata are sufficiently well represented in both training and tests sets. Large data sets are required to validate decision rules involving multiple genes and epistatic effects on distinct ethnic backgrounds. It is critical that delivery of personalized and predictive care be based on robust and accurate decision rules.

17.4.2 Genetic Signatures That Identify Increased Risk of Adverse Effects

The above observations based on the application of supervised machine learning approaches can be extended by the use of unsupervised clustering techniques in order to detect patterns and strata in data, and to provide additional support for conclusions derived using decision trees. In a recent prospective study, we evaluated the effect of a panel of variants in candidate genes on opioid-related respiratory depression in 347 children following tonsillectomy (Biesiada et al. 2014). Using unsupervised hierarchical clustering and a combination of candidate genotypes and clinical variables, we identified several distinct clusters of patients with high risk (36–38 %) and low risk (10–17 %) of respiratory depression; with the relative risk of respiratory depression for high *versus* low risk clusters of up to 3.8.

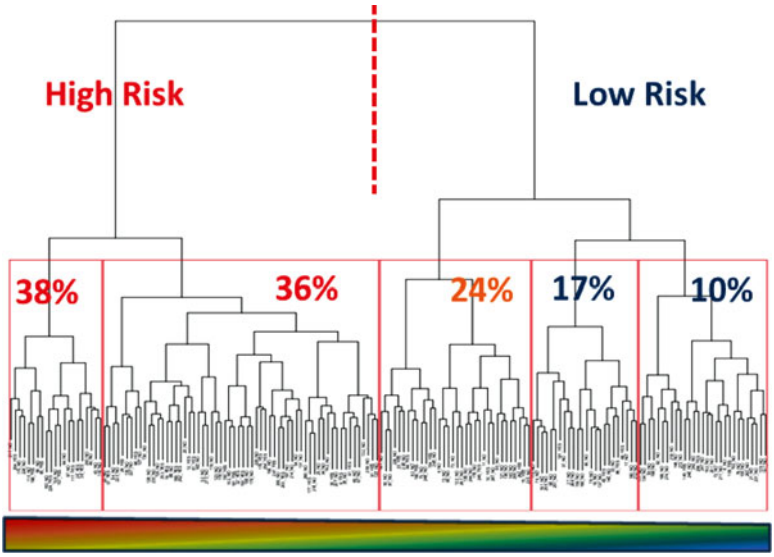


Fig. 17.8 Genetic signatures identify high vs. low risk subtypes among patients requiring post-surgical intervention: clustering of tonsillectomy patients using a set of candidate SNPs and clinical variables as features, and Hamming distance to define similarity (Note that two main clusters can be identified, which can be subsequently divided into five distinct sub-clusters with progressively increasing risk of RD (percent of RD cases in each cluster is shown))

As can be seen from Fig. 17.8, five clusters of patients that were characterized by different risk of RD can be distinguished using several SNPs (in genes discussed above) in conjunction with clinical and demographic variables to account for their confounding effects. These distinct clusters of patients are associated with statistically significant differences in either enrichment or under-representation of RD cases. Specifically, the frequency of RD cases increases gradually from about 10 % for the 'low risk cluster' to about 40 % for the 'high risk cluster'. Consequently, the relative risk of RD for the highest *versus* lowest RD risk clusters is about 3.8 ($p = 0.04$). Merging each of the 2 extreme clusters results in 3 clusters associated with low, intermediate and high risk of RD, while increasing the number of RD cases in low (RD frequency of 13.2 %) and high risk (RD frequency of 36.2 %) clusters. The relative risk between these two clusters is somewhat lower (2.7), but statistically significant ($p = 0.003$).

All non-redundant SNPs were included in the analysis at this stage, while SNP genotypes were encoded using two binary variables (see (Biesiada et al. 2014) for details). In addition to genetic variants, OSA, race, sex and morphine dose were also included for clustering to test for their effect on implicit stratification with respect to RD risk. Though these clinical or demographic variables were used in the analysis, they did not show an obvious association with the clusters identified, with the exception of African-American (AA) ancestry.

Children of AA ancestry predominantly (33 out of 39) clustered as a sub-cluster of the large high risk cluster with 36 % RD frequency among the total of 62 patients in this cluster. The lowest risk cluster (with just 10 % of RD cases) consisted of

almost exclusively white children (28 out of 29 in the cluster), whereas the highest RD risk cluster (with RD frequency of 38 %) consisted of 18 Caucasian and 3 non-Caucasian.

The fraction of patients who obtained the highest dose of morphine (>0.3 mg/kg) was actually somewhat lower for the highest risk cluster (28 %) in comparison with the lowest risk cluster (25 %). The fraction of OSA cases and ratio between males and females were also similar in these two extreme clusters. Thus, clearly factors other than race or morphine dose differentiate between high and low RD risk clusters.

The role of genetic factors in risk of RD is further supported by qualitatively similar clustering pattern obtained using just SNP genotypes (data not shown). In order to better elucidate the role of race in RD risk, clustering analysis was also performed separately for the subset of children of Caucasian descent only. In the latter case, using just SNP genotypes without any clinical variables, four clusters with increasing fraction of RD cases can be distinguished. Exclusion of patients of AA ancestry dissolves the racially mixed second high risk cluster of Fig. 17.8, while the highest risk Caucasian only cluster largely overlaps with the highest risk cluster observed using all individuals (and in conjunction with clinical and demographic variables) used for Fig. 17.8. Interestingly, an even higher cohesion and stronger association with RD is observed for the highest risk cluster, when limiting the analysis to Caucasians only, while using SNP genotypes only (50 % vs. 38 % of RD cases, respectively). Thus, race and ancestry based stratification plays an important role in the assessment and prediction of the risk of adverse opioid effects.

Consequently, the risk of RD could be identified using cluster assignment based on the relevant SNPs, with clinical and demographic variables contributing to more robust predictions. This approach resembles 'gene signatures'-based clustering of patients for cancer subtype classification (Miller et al. 2005). Genetic risk signatures, along with clinical risk factors, effectively identify children at higher and lower risks of opioid induced respiratory depression. Thus, genetic signatures of respiratory depression offer strategies for improved clinical decision support to guide clinicians to balance the risks of opioid adverse effects with analgesia.

17.5 Summing Up

Biomedical research is increasingly data driven and characterized by an exponentially growing body of sequence and other data. This challenge has motivated the development of platforms and algorithms to mine and analyze biomedical data with the goal of enabling clinical applications. In particular, translational and clinical studies rely on the ability to combine sequence and other molecular data with clinical and demographic information via clinical information systems. Given that sequence information from individual genomes is becoming ubiquitous, the challenge of interpreting and using these huge data sets in the context of the individual

genetic variability becomes an even more pressing problem. A critical advantage of *in silico* genomic and biomedical studies is that computational methods can take advantage of current computing power to enable complex analyses, which shed light on the intricate networks of molecular interactions, hubs and pathways that underlie observed phenotypes. Understanding these networks can lead to better therapeutic interventions as discussed in Chaps. 16 and 20.

Interpreting the vast (and rapidly growing) amounts of data leads to problems of size and complexity that cannot be solved in routine ways; new algorithms and tools are required. Additionally, the simultaneous consideration of many different data types, such as expression profiles and polymorphisms, clinical and demographic data etc., poses significant challenges for these methods. It also requires that a new generation of highly integrated and versatile informatics systems be developed that combine databases with analytical and visualization tools.

This chapter provides several case studies to illustrate the use of variant-based stratification strategies in the context of pediatric disease. Specific applications are discussed in this chapter to illustrate how sequencing and clinical data stratification can be applied in the context of translational and basic research. The first application deals with the prediction of HLA alleles from SNP data. Many specific variants (alleles) of classical HLA genes have been implicated as risk or protective factors in pediatric autoimmune disorders, including juvenile rheumatoid arthritis, type 1 diabetes and celiac disease. Typing classical HLA genes is also commonly used to match transplant donors and recipients. Integrated informatics solutions for storing, processing and analyzing SNP and other sequencing data from multiple clinical studies can greatly facilitate efforts to extend and improve prediction of HLA alleles. Such databases and tools provide consistent quality control criteria and protocols, as well as access to data potentially pertaining to cohorts of different ancestries that can be used to assess, refine and re-train current methods.

Postoperative care and pain management in children provides another example of application of variant data. The use of genetic polymorphisms in conjunctions with clinical data can guide therapeutic decisions that personalize and improve care. However, informatics and other challenges must be addressed by researchers and clinicians aiming to optimize such personalized interventions. Multi-center studies are generally necessary in pediatrics to accrue sufficient numbers of patients to assure adequate statistical power to reach significant, reliable conclusions. Such studies pose complex challenges in data capture, warehousing, sharing, and analysis. Data systems of different institutions participating in a multi-center study, especially their electronic health records, frequently lack semantic and syntactic interoperability, which adds to the difficulties. These issues are discussed in greater detail in Chaps. 3 and 6. We believe that approaches similar to those outlined in this chapter can be applied in other translational studies that aim to develop decision rules and to implement evidence based decision support systems for personalized therapeutic interventions.

References

Aureli A, et al. Identification of a novel HLA-B allele, HLA-B*3580, with possible implication in transplantation and CTL response. Tissue Antigens. 2008;71(1):90–1.

Biesiada J, et al. Genetic risk signatures of opioid-induced respiratory depression following pediatric tonsillectomy. Pharmacogenomics. 2014;15(14):1749–62.

Caldas JC, et al. General anesthesia, surgery and hospitalization in children and their effects upon cognitive, academic, emotional and sociobehavioral development – a review. Paediatr Anaesth. 2004;14(11):910–15.

Cepeda MS, et al. Side effects of opioids during short-term administration: effect of age, gender, and race. Clin Pharmacol Ther. 2003;74(2):102–12.

Chan IS, Ginsburg GS. Personalized medicine: progress and promise. Annu Rev Genomics Hum Genet. 2011;12:217–44.

Chou CK, et al. Human insulin receptors mutated at the ATP-binding site lack protein tyrosine kinase activity and fail to mediate post receptor effects of insulin. J Biol Chem. 1987;262(4):1942–7.

Chou W-Y, et al. Human opioid receptor A119G polymorphism affects intravenous patient-controlled analgesia morphine consumption after total abdominal hysterectomy. Anesthesiology. 2006a;105(2):334–7.

Chou W-Y, et al. Association of mu-opioid receptor gene polymorphism (A119G) with variations in morphine consumption for analgesia after total knee arthroplasty. Acta Anaesthesiol Scand. 2006b;50(7):787–92.

Coller JK, et al. ABCB1 genetic variability and methadone dosage requirements in opioid-dependent individuals. Clin Pharmacol Ther. 2006;80(6):682–90.

Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.

de Bakker PI, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 2006;38(10):1166–72.

Diatchenko L, et al. Genetic basis for individual variations in pain perception and the development of a chronic pain condition. Hum Mol Genet. 2005;14(1):135–43.

Diatchenko L, et al. Catechol-O-methyltransferase gene polymorphisms are associated with multiple pain-evoking stimuli. Pain. 2006;125(3):216–24.

Diatchenko L, et al. Genetic architecture of human pain perception. Trends Genet. 2007;23(12):605–13.

Dilthey AT, et al. HLA*IMP–an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics. 2011;27(7):968–72.

Dinov I, et al. Efficient, distributed and interactive neuroimaging data analysis using the LONI pipeline. Front Neuroinform. 2009;3(22):1–10.

Duedahl TH, Hansen EH. A qualitative systematic review of morphine treatment in children with postoperative pain. Paediatr Anaesth. 2007;17(8):756–74.

Esclamado RM, et al. Perioperative complications and risk factors in the surgical treatment of obstructive sleep apnea syndrome. Laryngoscope. 1989;99(11):1125–9.

Hastie T, et al. The elements of statistical learning. 2nd ed. New York: Springer; 2009.

Holmes M, et al. Fulfilling the promise of personalized medicine? Systematic review and field synopsis of pharmacogenetic studies. PLoS One. 2009;4(12):e7960.

Hothorn T. Unbiased recursive partioning: a conditional inference framework. J Comput Graph Stat. 2010;15(3):651–74.

Howie BN, et al. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529.

Kent WJ, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006.

Klepstad P, et al. The 119 A>G polymorphism in the human mu-opioid receptor gene may increase morphine requirements in patients with pain caused by malignant disease. Acta Anaesthesiol Scand. 2004;48(10):1232–9.

LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucl Acids Res. 2009;37(13):4191–3.
Lechler R, Warrens A. HLA in health and disease. San Diego: Academic; 2000.
Leslie S, et al. A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet. 2008;82(1):48–56.
Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754–60.
Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165(4):2213–33.
Li M, et al. Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. Am J Hum Genet. 2005;76(6):934–49.
Manichaikul A, et al. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867–73.
Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11:499–511.
Marchini J, et al. A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet. 2006;78(3):437–50.
Marchini J, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007;39(7):906–13.
Marsh SGE, et al. Nomenclature for factors of the HLA system. Tissue Antigens. 2010a;75(4):291–455.
Marsh SGE, et al. An update to HLA nomenclature. Bone Marrow Transplant. 2010b;45(5):846–8.
McKenna A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
Miller LD, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A. 2005;102(38):13550–5.
Nackley AG, Diatchenko L. Assessing potential functionality of catechol-O-methyltransferase (COMT) polymorphisms associated with pain sensitivity and temporomandibular joint disorders. Methods Mol Biol. 2010;617:375–93.
Nackley AG, et al. Catechol-O-methyltransferase inhibition increases pain sensitivity through activation of both beta2- and beta3-adrenergic receptors. Pain. 2007;128(3):199–208.
Oertel BG, et al. The mu-opioid receptor gene polymorphism 119A>G depletes alfentanil-induced analgesia and protects against respiratory depression in homozygous carriers. Pharmacogenet Genomics. 2006;16(9):625–36.
Orozco G, et al. Auto-antibodies, HLA and PTPN22: susceptibility markers for rheumatoid arthritis. Rheumatology (Oxford). 2008;47(2):138–41.
Pappas D, et al. Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest, submitted. 2016.
Patel ZH, et al. The struggle to find reliable results in exome sequencing data: filtering out Mendelian errors. Front Genetics. 2014;5(16).
Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9.
Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
Quinlan JR. Induction of decision trees. Machine Learning 1986;1(1):81–106.
Rakvåg TT, et al. The Val158Met polymorphism of the human catechol-O-methyltransferase (COMT) gene may influence morphine requirements in cancer pain patients. Pain. 2005;116(1–2):73–8.
Rakvåg TT, et al. Genetic variation in the catechol-O-methyltransferase (COMT) gene and morphine requirements in cancer patients with pain. Mol Pain. 2008;4:64.

Ritchie ME, et al. R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. Bioinformatics. 2009;25(19):2621–3.
Robinson J, et al. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res. 2003;31(1):311–14.
Robinson J, et al. The IMGT/HLA database. Nucleic Acids Res. 2009;37(Database issue):D1013–17.
Sadhasivam S, et al. Race and unequal burden of perioperative pain and opioid related adverse effects in children. Pediatrics. 2012;129(5):832–8.
Sampaio-Barros PD, et al. Frequency of HLA-B27 and its alleles in patients with Reiter syndrome: comparison with the frequency in other spondyloarthropathies and a healthy control population. Rheumatol Int. 2008;28(5):483–6.
Schaaf CP, et al. Copy number and SNP arrays in clinical diagnostics. Annu Rev Genomics Hum Genet. 2011;12:25–51.
Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78(4):629–44.
Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet. 2003;73(5):1162–9.
Stephens M, et al. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 2001;68(4):978–89.
The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
The International HapMap Consortium. The international HapMap project. Nature. 2003;426:789–96.
Therneu TM, Atkinson B. rpart: Recursive partitioning. 2009. http://CRAN.R-project.org/package=rpart.
Wang K, et al. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78.
Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005.

Chapter 18
Application of Genomics to the Study of Human Growth Disorders

Michael H. Guo and Andrew Dauber

Abstract Modern genomic approaches have helped elucidate the genetic basis of many diseases. Here, we use growth disorders as a paradigmatic example of how modern genomics can transform our understanding of biology and disease. Disorders of growth—either short stature or overgrowth—are frequently associated with genetic mutations. However, as the large majority of patients with growth disorders lack a molecular diagnosis, genetic studies have the potential to transform our ability to understand, diagnose, and ultimately treat these disorders. In this chapter, we discuss how modern genomic methods such as next generation sequencing have led to the discovery of novel genes and helped expand the phenotypic spectrum associated with known genes. Additionally, we discuss the translation of these technologies into the clinic as diagnostic tools for patients with growth disorders. We also discuss the application of SNP genotyping to find genetic variants and genes associated with human stature in the general population and how these studies have generated insights into the biology of human growth. Throughout, we highlight some of the challenges with these technologies in the application to the understanding of human disease and biology.

Keywords Genomics • Whole exome sequencing • Whole genome sequencing • Chromosomal microarray • SNP genotyping

M.H. Guo, Ph.D. (✉)
University of Florida College of Medicine, Gainesville, FL, USA

Department of Genetics, Harvard Medical School, Boston, MA, USA, 02115
e-mail: michael_guo@hms.harvard.edu

A. Dauber, M.D., M.Msc
Departments of Pediatrics, Cincinnati Center For Growth Disorders, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine,
3333 Burnet Avenue, Cincinnati 45229, OH, USA
e-mail: andrew.dauber@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_18

18.1 Introduction

Historically, the mapping of disease genes has largely relied on linkage analyses in large families with multiple affected individuals. Linkage analyses, however, provide limited resolution in attempts to find a disease-causing gene (Altshuler et al. 2008; Pulst 1999) and require large carefully ascertained pedigrees. These pedigrees are often difficult to collect, if available at all. Once linkage analysis is performed, linked regions are often large and require much more painstaking work to refine down to the causal gene within the linkage region. Thus, even by the late 1990s, only a handful of genes had been associated with human disease.

Over the past decade, next generation sequencing and other high throughput technologies have essentially supplanted linkage as the primary mode of Mendelian gene discovery. These technologies have spurred the rapid discovery of genes associated with human disease, including associating mutations in hundreds of genes with various Mendelian disorders, as well as tens of thousands of genomic loci with complex diseases and traits (Chong et al. 2015; Koboldt et al. 2013; Stranger et al. 2011; Visscher et al. 2012). Technologies such as chromosomal microarrays, next generation sequencing (whole genome and whole exome) and SNP genotyping allow for comprehensive and accurate genotyping of various forms of genetic variation. As these technologies are typically applied genome-wide, they have the added advantage of being limited by relatively few *a priori* assumptions or hypotheses.

Beyond its role in gene discovery, genetics is also an incredibly powerful tool for understanding biology and disease (Hirschhorn 2009). For diseases with a genetic basis, the genetic mutation represents the most proximal factor contributing to disease, and thus identification of the genetic mutation can identify the causal (rather than correlated) factors for disease. The results of genetic studies have helped uncover genes and pathways underlying the pathophysiology of disease, and many of these insights have been completely unsuspected (Hirschhorn 2009; Visscher et al. 2012). For example, a genome wide association scan identified a variant in the complement factor H gene that strongly increased risk for age-related macular degeneration (AMD) (Klein et al. 2005).

In this chapter, we review the application of modern genomic technologies to elucidating the genetic basis of human disease. We focus primarily on its application in the context of short stature, a clinically relevant phenotype for which these technologies have offered immeasurable insight. We will outline how next generation sequencing has catalyzed the discovery of novel disease genes, helped expand phenotypes associated with known disease entities, helped us understand the spectrum of mutations that can cause a given phenotype, and ultimately how it is being applied in the clinic. We also outline how chromosomal microarrays are helping to understand the role of copy number variation in short stature. We close by discussing the insights that genome wide association studies have generated into the biology of human growth. Throughout, we highlight some of the limitations of these technologies.

18.2 Background to Growth and Growth Disorders

Growth disorders, in particular short stature, are a common presentation to pediatric endocrinology clinics. Despite tremendous improvements in our understanding of human growth and growth disorders, our ability to diagnose, manage, and treat short stature patients is quite poor. Depending on the setting, only 1–40 % of short stature patients undergoing a standard medical evaluation have an identifiable cause for their short stature (Ahmed et al. 1993; Green and MacFarlane 1983; Grimberg et al. 2005; Grote et al. 2008; Lindsay et al. 1994; Sisley et al. 2013; Voss et al. 1992). Thus, there is a tremendous need to improve our ability to understand the genetic and molecular causes of growth disorders.

Many disparate biological processes influence growth, including hormonal signaling, growth factor signaling, basic cellular processes such as DNA replication and cell division, as well as intercellular interactions. Here we highlight a few of the major pathways and process to familiarize the reader with some of the important components controlling human growth and how defects in these components can lead to disease. The primary hormonal axis controlling growth is the growth hormone/insulin-like growth factor axis (David et al. 2011). Deficiency of hormones in the axis, resistance at the hormone receptors, or defects in hormone binding proteins can all result in growth disorders. Many signaling molecules converge at the growth plate of long bones, where the differentiation and maturation of chondrocytes and elongation of the growth plate during development drives human height (Baron et al. 2015). Perturbations to chondrocyte development or function, intercellular signaling pathways in the growth plate, or extracellular matrix proteins that guide growth plate formation can all result in growth disorders. Interestingly, many core cellular processes are also involved in growth. Defects in centrosome formation, DNA repair, histone modifications, and cell proliferation have all been associated with human growth disorders (Klingseisen and Jackson 2011).

In this chapter, we focus on short stature. While short stature disorders can be classified in any number of ways, they can be broadly classified as: (1) disorders presenting with prenatal onset growth retardation, (2) disorders of the growth hormone/insulin-like growth factor axis, (3) skeletal dysplasias (defects in skeletal formation), and (4) idiopathic short stature (short stature of unknown etiology) (Dauber et al. 2014b). To date, there are over 400 Mendelian growth disorders identified, and the functions of the proteins perturbed by the genes for these disorders span a dizzying array of intracellular and intercellular processes (de Bruin and Dauber 2016).

18.3 Human Height as a Model Genetic Trait

The genetics of human stature is also an intense area of study not only because it can provide valuable insights into the biology of human growth, but also because it is considered a model genetic trait. Human height has been studied for centuries and

continues to be a subject of intense speculation into its genetic architecture (the spectrum of genetic mutations and their relative contributions to traits and diseases) (Hirschhorn and Lettre 2009). Height is a highly heritable trait, with over 80 % of variation in height attributable to genetic variation (Silventoinen et al. 2003; Yang et al. 2015). It is also clearly influenced by both common and rare genetic variation. Genome wide association studies (GWAS) over the last decade have helped elucidate the role of common genetic variants (those with minor allele frequency above 5 %), each having small effects on height, but which together wield a sizeable influence on height (Wood et al. 2014; Yang et al. 2015). Rare genetic variants are also clearly important, as highlighted by the multitude of Mendelian growth disorders that are caused by rare mutations of large phenotypic effect (de Bruin and Dauber 2016). Interestingly, for individuals at the short extreme of the height distribution, their height is shorter than would be predicted based on the cumulative predicted effect of their common genetic variants (Chan et al. 2011). This suggests that rare variant(s) with larger phenotypic effect(s) are driving their short stature. Nonetheless, the exact relative contribution of common and rare alleles is as of yet unknown, but is likely to be unraveled over the coming years with improved technologies and ever-expanding sample sizes.

18.4 Application of Next Generation Sequencing to Human Growth Disorders

Here, we provide a brief background to next generation sequencing to acclimate the reader to these technologies, although we refer readers to excellent reviews that explain the technologies in greater detail. Next generation sequencing refers to high throughput sequencing methods, which allow for parallelization of the sequencing method in contrast to traditional Sanger sequencing. Many different technologies for NGS exist, including sequencing by synthesis (Illumina), sequencing by ligation (SOLiD), and single molecule real time sequencing (PacBio) (Mardis 2013). These technologies generate sequencing "reads" which are simply readouts of the sequence of a segment of an individual's DNA. These reads vary in length depending on the technology. Once the sequence of each read is determined, the reads can then be "aligned" to a reference genome and various variant-calling algorithms can help determine differences from the reference genome (genetic variants or mutations) (DePristo et al. 2011; Liu et al. 2013; Mielczarek and Szyda 2016; Nelson et al. 2015). NGS allows for high coverage, which means that at any given site in the genome, many reads are generated, allowing for higher confidence in the detection of differences as compared to the reference genome sequence (Mielczarek and Szyda 2016; Robinson and Storey 2014).

NGS can be whole genome, whole exome, or targeted sequencing. In whole genome sequence, all three billion bases of the human genome are sequenced, allowing for comprehensive coverage of all forms of genetic variation. In whole

exome sequencing, custom capture probes allow for isolation and sequencing of just the protein-coding regions (~1.5 % of the genome) (Ng et al. 2009). In contrast to whole genome, whole exome typically allows for higher coverage of the protein-coding regions and is significantly cheaper. In general, with either exome or genome sequencing, the protein-coding portions of the genome are the only portions that are analyzed as they are much more interpretable than the noncoding portions of the genome. However, exome sequencing may often miss areas of protein-coding regions that are poorly captured by the probes (Belkadi et al. 2015; Meynert et al. 2014). Finally, researchers can also design custom probes to target genes or regions of interest. This allows for even greater coverage of regions of interest and potential cost-savings as compared to whole exome or genome.

While NGS technologies have transformed our ability to read the human genome, there are still many limitations to the technology. The technology introduces many errors and artifacts in the process of generating sequencing reads (Mielczarek and Szyda 2016; Nielsen et al. 2011). Identifying true genetic variants and differentiating them from the numerous sequencing artifacts remains a formidable challenge (DePristo et al. 2011; Mielczarek and Szyda 2016; Nielsen et al. 2011). Certain regions of the genome, in particular low-complexity regions, are difficult to sequence and align to the reference genome (Weisenfeld et al. 2014). Thus, while the technology has fundamentally changed our ability to identify mutations, technological developments in the next few years will allow for more accurate and cheaper sequencing.

Once the genetic variants of a given individual have been determined, making sense of this data is even more challenging. Most genetic variants in the genome are benign and for the most part, it is very difficult to differentiate the few truly pathogenic variants from the vast background of benign variations in the human genome. Generally, for Mendelian disorders, only rare variants are considered, under the assumption that more common variants are likely to be eliminated by natural selection if they have a large effect on disease risk (Goldstein et al. 2013). Various algorithms also exist to try to predict the likely deleterious effects of a protein-coding variant, but these algorithms have imperfect sensitivity and specificity (Gnad et al. 2013; Hicks et al. 2011). The challenge of interpreting variants is compounded in the noncoding regions, where we still have very limited understanding of how sequence variants can perturb noncoding functional elements to influence gene expression.

18.4.1 Identification of Novel Disease Genes

One of the primary goals of genetics research is to identify novel genes underlying human disease. The typical approach for identifying novel disease genes is to collect a cohort of patients with a given phenotype, determine their genetic sequence, and identify genes which carry mutations in more patients than would be expected by chance based on some statistical measure. As generating sufficient statistical

signal is difficult given the sample sizes typically available for rare disorders, nominating a novel disease gene often relies on corroborating functional evidence. An observation of a mutation in a gene in a single patient or even a single family is generally insufficient to nominate a new disease gene as causal for the disorder. Here, we highlight our efforts that led to the discovery of three new genes associated with growth disorders. For all three of these cases, we incorporated evidence across patients and families with the same phenotype to increase our confidence in our assignments of novel disease genes.

Exome sequencing of a patient with severe short stature, developmental delay, microcephaly, and craniofacial and cardiac defects and her parents revealed a *de novo* missense mutation in *PUF60,* a gene known to function as an RNA splicing factor. Interestingly, acrofacial dystosis, Nager type (MIM 154400) and mandibulofacial dysostosis with microcephaly (MIM 610536), two syndromes with striking overlap with this patient's presentation, are also due to mutations in splicing factors *SF3B4* and *EFTUD2*, respectively. These factors were known to interact with PUF60 at the protein level (Hastings et al. 2007). However, this observation alone was not sufficient to implicate the *de novo* mutation in *PUF60* as causal. *In vitro* studies demonstrated that the patient's specific missense mutation led to alteration in splicing of known PUF60 targets, further supporting a pathogenic effect of this variant. Additionally, an international collaboration subsequently identified five patients with similar phenotypes who carried deletions that encompassed *PUF60* and a nearby gene, *SCRIB.* Zebrafish models of *PUF60* and *SCRIB* deficiency were generated which helped delineate which parts of the phenotype were due to *PUF60* deficiency versus loss of *SCRIB* function. The combination of evidence from the *de novo* mutation in *PUF60,* the additional five patients with deletions of *PUF60/SCRIB*, and extensive functional testing and animal models helped to establish *PUF60* as the causal gene for this patient's syndrome (Dauber et al. 2013a). This case demonstrates that implicating novel disease genes can be quite challenging and will typically necessitate several independent patients with mutations in the same gene, along with extensive functional work.

We performed exome sequencing in multiple members of two unrelated families who presented with growth retardation and marked elevation of total IGF-I and IGFBP-3 concentrations. Exome sequencing analysis revealed two different homozygous mutations in *PAPPA2* that were found to segregate with the phenotype in each family (Dauber et al. 2016). As there are relatively few rare protein-altering variants that are homozygous in two different members of a given family, we were able to rapidly filter down to a small set of variants for each family. Mutations in *PAPPA2* were the only segregating autosomal recessive variants shared between the families. Both mutations were predicted to be functionally deleterious for the PAPP-A2 protein, and *in vitro* analysis of IGFBP cleavage demonstrated that both mutations cause a complete absence of PAPP-A2 proteolytic activity. We then showed that the patients' serum had decreased levels of free IGF-I and decreased IGF bioactivity despite the marked elevation in total IGF-I levels. These patients are the first individuals to have short stature resulting from mutations in the IGF binding

proteins leading to decreased IGF-I bioavailability. They provide novel insights into the growth hormone/IGF signaling pathway.

In a set of three families with a rare phenotypic constellation of mild short stature, advanced bone age, and early growth cessation, we identified heterozygous variants in *ACAN* (Nilsson et al. 2014). The *ACAN* gene encodes aggrecan, a proteoglycan that is an important component of the extracellular matrix and has a known role in growth plate formation (Aspberg 2012). While variants in *ACAN* had previously been linked to severe dwarfism disorders that presented with joint defects and/or skeletal malformations (Gleghorn et al. 2005; Stattin et al. 2010; Tompson et al. 2009), our patients represented a novel phenotypic entity. The discovery of *ACAN* mutations in this disorder of mild short stature, advanced bone age, and early growth cessation relied on overlap in genetic variants across families. Following filtering for rare, protein-altering variants that segregated with the phenotype in each family, we found 60, 19, and 1 variant that met these criteria in the respective families. Remarkably, different variants in *ACAN* meeting these criteria were seen in each of the three families. In fact, *ACAN* is the only gene shared by any two of the three families. These variants were predicted to either be protein-truncating or were missense mutations in critical domains of the protein. The discovery of *ACAN* causing this rare constellation of features relied on evidence across families. The observation of all three families sharing mutations in the same gene is extremely unlikely by chance and allowed us to filter through the other variants that segregated with the phenotype but likely do not cause the disorder. Subsequent studies have identified additional families with this set of features who also carry rare protein-altering variants in *ACAN* (Quintos et al. 2015).

18.4.2 Expansion of Phenotypes

Comprehensive examination of genetic variation in patients has also helped expand the phenotypic spectrum associated with known disease genes. With next generation sequencing and the ability to comprehensively and relatively cheaply ascertain all genetic variation, genetics is moving toward a genotype-first approach. Traditionally, patients with a specific disorder were recruited based on their phenotype and then genotyped to identify genetic changes that were present in more patients than expected by chance. Under a genotype-first approach, the genotype of the patient serves as a starting point, and the phenotypic features of the patients carrying a given mutation or mutations in a given gene are compared. While the distinction may at first seem subtle, the genotype-first approach has a tremendous benefit in that it allows us to expand the phenotypic spectrum associated with genotypes. This approach has led to the observation that many genetic mutations previously known to cause a specific phenotype actually can be associated with a much wider phenotypic spectrum or only rarely causes disease (incomplete penetrance) (Minikel et al. 2016). Although truly genotype-first genetic studies are just ramping up due to the extensive sample sizes needed, here, we use several examples from our

research program to illustrate how ascertaining patients on a relatively nonspecific finding of short stature and comprehensively examining their genetic variation can help widen the phenotypic spectrum associated with diseases.

Sequencing of an affected sib-pair with short stature and developmental delay identified compound heterozygous mutations in *SLC35C1* (Dauber et al. 2014a). Mutations in *SLC35C1* had previously been found to cause Leukocyte Adhesion Deficiency type II (LADII) (MIM 266265), a hereditary disorder of fucosylation (addition of carbohydrate groups to glycoproteins or glycolipids) that results in inability of granulocytes to bind selectins, manifesting as marked leukocytosis (increase in white blood cell count) and recurrent bacterial infections as well as short stature and developmental delay (Etzioni et al. 1992). However, extensive functional testing revealed that the two siblings had only partial leukocyte adhesion deficiency and did not have leukocytosis or recurrent infection. The only apparent phenotypic presentation of the *SLC35C1* variants was the short stature and developmental delay. This expanded the phenotype associated with *SLC35C1* mutations and suggested an unexpected role of fucosylation in growth. The molecular diagnosis in this case would not likely have been made with traditional targeted genetic approaches, since the patient would not be suspected to have LADII. By broadly evaluating genetic variation at a large number of genes through targeted sequencing of over 1000 genes (see below), we were able to search for potentially pathogenic variants and then determine which of these were likely to cause the phenotype. Finally, this diagnosis also has potentially important treatment ramifications had it been made earlier in life, as fucose supplementation has been shown to improve the hematological outcomes in patients with LADII with the effects on cognition and growth still unknown (Marquardt et al. 1999).

In an affected pair of siblings with a syndrome of severe postnatal growth retardation, gonadal failure, microcephaly, and early-onset metabolic syndrome, exome sequencing revealed homozygous variants in *XRCC4* (de Bruin et al. 2015). While homozygous mutations in *XRCC4* had previously been seen in a single individual with short stature and microcephaly (Shaheen et al. 2014), this report of affected siblings with *XRCC4* mutations expanded the phenotypic spectrum to include gonadal failure and early-onset metabolic syndrome (Shaheen et al. 2014). This case also illustrates how the approach of sequencing affected sib-pairs can be very powerful. Each individual generally only carries a few rare, recessive, protein-altering variants. As siblings are only expected to share ¼ of recessive mutations, the number of candidates can be quite small based on genetics alone. Another paper published contemporaneously also identified bi-allelic (on both copies of the gene) mutations in *XRCC4* in five families with short stature and microcephaly (Murray et al. 2015). The association of *XRCC4* mutations with growth retardation is only one of many examples of how genes in core cellular processes can cause growth defects. Interestingly however, patients with mutations in *XRCC4* did not develop immune defects (de Bruin et al. 2015; Murray et al. 2015). This is counterintuitive given the critical role of *XRCC4* in non-homologous end joining, which is believed to be necessary for V(D)J recombination in the formation of immune molecules (de

Villartay 2015). Thus, human patients present a unique opportunity to better understand biology.

In an individual patient who was ascertained on idiopathic short stature (ISS), we found compound heterozygous mutations in *B4GALT7* (Guo et al. 2013), a gene previously known to cause the progeroid form of Ehlers Danlos Syndrome (MIM 130070) (Faiyaz-Ul-Haque et al. 2004; Kresse et al. 1987), but had not been linked to ISS. Upon further phenotypic workup of the patient, he was found to have some features consistent with the progeroid form of EDS, including hyperflexible joints, hyperextensible skin, and congenital malformation of the elbow joint. However, he did not have progeroid facial features, the phenotypic feature for which the disorder was named. Upon literature review of the previously known 3 cases of the progeroid form of EDS, it was found that progeroid facies are not a consistent feature of patients carrying *B4GALT7* (Faiyaz-Ul-Haque et al. 2004; Kresse et al. 1987). Interestingly, the previously known patients with *B4GALT7* mutations did have short stature. Thus, we were able to redefine the progeroid form of Ehlers Danlos syndrome to include short stature and demonstrate that progeroid features are not a phenotypic feature of patients with *B4GALT7* mutations (Guo et al. 2013).

In another patient presenting with ISS, we identified a known mutation in the gene *FAM111A* that causes Kenney-Caffey syndrome (Guo et al. 2014). While the patient had all the other characteristic features of KCS, this patient did not have infantile hypocalcemia, the cardinal feature of KCS (Unger et al. 2013). As the patient was missing this key feature of KCS, this diagnosis—which is already very rare—would have been highly unlikely to have been considered using traditional genetic approaches, but was made possible by unbiased genotyping in the form of whole exome sequencing.

Our studies comprehensively evaluating genetic variation in patients ascertained on a relatively wide and nonspecific phenotype of short stature is a step toward a genotype-first approach and illustrates how this approach can be illuminating in expanding the phenotypic spectrum associated with disease. We were able to find many patients with mutations in known disease genes that did not fit the phenotypic definitions classically associated with mutations in those genes. In the future, with expanded resources, population-based ascertainment and genotyping will likely continue to greatly expand the phenotypic spectrum of these genes.

18.4.3 Understanding the Spectrum of Mutations That Cause Disease

While there are more than 400 genes known to cause Mendelian growth disorders, the mutations within these genes are not well catalogued. Even within Mendelian disease genes, most genetic mutations are benign. Thus, understanding the mutations that can contribute to disease is important in helping to interpret genetic variants in these genes and can aid in future diagnoses. However, doing so in a high

throughput fashion is challenging given the current costs of sequencing, especially when applied to a large number of genes.

In order to identify potential disease-causing mutations in genes known to cause short stature, we performed a targeted sequencing experiment to sequence 1077 genes in 192 short stature patients along with 192 controls (Wang et al. 2013). These 1077 genes were selected based on their status as a Mendelian growth disorder gene, previous links to growth plate biology, or being found in a GWAS locus for height (see below). In order to overcome the cost limitations of current technology, we undertook a pooled sequencing approach which allowed for highly accurate sequencing of all 1077 genes, while saving significantly on sequencing costs (Golan et al. 2012). The study revealed 4928 genetic variants in 1077 genes that were present in cases but not in controls, of which 1349 had not been seen before in large reference databases. Among the many interesting findings from the study, the analyses identified three individuals with known pathogenic variants in *PTPN11* causing undiagnosed Noonan syndrome. The analyses also revealed 9 rare potentially nonsynonymous variants in *IGF1R*. Interestingly, most of these variants did not segregate with the phenotype in the families from which the variant was found. This could suggest that the variants are either incompletely penetrant (only causing disease some of the time) or that they are not pathogenic. One of these variants in *IGF1R* was a novel frameshift variant, which was functionally tested and deemed to be likely pathogenic. Another known pathogenic variant in *IGF1R* was found in a control subject questioning the previous assertion that this is a pathogenic variant. Together, these data helped generate an extensive catalogue of variants in relevant genes in short stature patients and will greatly aid in interpreting which variants are likely pathogenic or benign.

The pooled sequencing analysis also revealed seven protein-altering variants in *NPR2* that were found exclusively in the short stature patients (Wang et al. 2013, 2015). Bi-allelic variants in *NPR2* had previously been demonstrated to cause a rare dwarfing skeletal dysplasia: acromesomelic dysplasia, Maroteaux type (AMDM) (MIM 602875). These variants were found in the heterozygous state in the short stature patients, suggesting that heterozygous variants in *NPR2* can contribute to idiopathic short stature, as had been previously suggested (Bartels et al. 2004; Olney et al. 2006). Several of these variants segregated with the short stature within the respective families or were found to be *de novo*. Screening of additional cohorts ascertained on population height extremes (short and tall stature) revealed nine additional rare nonsynonymous variants, eight of which were found in short stature patients and one of which was found in a tall stature patient. These variants were confirmed for functional effect, including demonstration of loss of function for variants found in short stature patients and gain of function for the variant from the tall stature patient (Wang et al. 2015). These studies greatly expanded the spectrum of variants in *NPR2* associated with short stature and demonstrated that heterozygous variants in the gene account for ~2% of patients presenting with idiopathic short stature. The observation of a tall stature patient with a gain of function variant further suggested that both loss and gain of function in *NPR2* can result in opposite changes in height.

18.4.4 Utility of Exome Sequencing in Diagnosing Short Stature in Clinical Settings

Next generation sequencing has also been moving from research laboratories to the clinic, where it has the potential to supplant traditional clinical genetics approaches of targeted sequencing of disease genes suspected based on a patient's phenotype. It provides the potential to provide genetic diagnoses for rare disorders, which can guide treatment and management of patients. However, as the technology is relatively new and no long-term follow-up studies have been performed, the efficacy of the approach has not been rigorously evaluated.

To evaluate the efficacy of NGS in providing diagnoses for patients, we performed exome sequencing in 14 patients with ISS and their unaffected family members (Guo et al. 2014). Patients all had extensive workup that did not reveal a cause of their short stature and did not have syndromic features suggestive of a specific genetic syndrome. Following filtering for Mendelian patterns of segregation (autosomal recessive, compound heterozygous, X-linked recessive, or *de novo*), we identified a genetic cause for 5 of the 14 patients (36 % diagnostic yield). Our diagnostic yield was similar to the yields reported for exome sequencing as applied to other rare Mendelian disorders (Lee et al. 2014a; Yang et al. 2014). These results suggested that exome sequencing might be effective in the diagnosis of patients with ISS, which is a relatively mild phenotype as compared to the severe phenotypes typically analyzed by molecular diagnostic clinics.

However, 9 of the 14 patients did not receive a genetic diagnosis. As we had hypothesized that these patients' short stature is due to monogenic highly rare penetrant genetic variants following a classic pattern of Mendelian inheritance, patients not following these assumptions are likely to evade diagnosis due to several different reasons. First, our approach was predicated on the assumption that mutations in a single gene resulted in the ISS, but there is the possibility of genetic variants in multiple genes contributing to the short stature (i.e. oligo- or polygenicity). Second, more common mutations (>1 % minor allele frequency in reference databases) could also cause the short stature, although these variants are unlikely to have a large effect on stature or are unlikely to be highly penetrant. Additionally, synonymous (mutations that do not change the protein sequence) changes were not considered as they are presently not amenable to interpretation, although it is well known that synonymous changes can perturb gene expression and function. Non-coding variants, epigenetic changes, and somatic changes (i.e. non-germline) are not captured well by exome sequencing. A final challenge is that since we relied on segregation in Mendelian patterns, we relied on correct specification of other family members, as being affected or unaffected, and the mutations being fully penetrant. Misspecification of phenotypic status in other family members can misguide the segregation analyses and prevent the identification of a causal genetic variant. Similarly, incomplete penetrance can result in the causal mutation being present in an unaffected family member, preventing the mutation from demonstrating Mendelian segregation.

Interpreting genetic variation can be quite challenging, as most genetic variants are likely benign and for most variants, we have limited ability to resolve whether they are pathogenic or benign. For example, in our paper, one patient with ISS was

found to carry compound heterozygous variants in *COL2A1,* which is known to cause a variety of skeletal dysplasia and short stature syndromes (Kannu et al. 2012). However, the patient does not have any known skeletal anomalies, and it is unknown if this patient's variants in *COL2A1* could cause ISS in the absence of skeletal defects.

We also identified two patients with 3-M syndrome. While this disorder was believed to be rare (Hanson et al. 2011), our identification of two patients (out of 14 analyzed) suggested that it might actually be underdiagnosed. These two patients had both undergone extensive clinical evaluation that did not identify a cause for the ISS. The diagnosis of 3-M syndrome has important treatment implications, as the 3-M syndrome is often associated with hypergonadotropic hypogonadism (Dauber et al. 2013b). The identification of the association between 3-M and hypergonadotropic hypogonadism opens up the path for patients to potentially undergo fertility preservation therapy.

Whole exome sequencing might also be a cost-effective modality to providing a diagnosis for patients with short stature. One of the patients who was diagnosed with 3-M syndrome had undergone nearly $8000 in various clinical genetic tests and targeted sequencing (Dauber et al. 2013b). At the time, clinical exome sequencing also cost approximately $8000, but the prices for exome sequencing are dropping rapidly. This suggests that exome sequencing has the potential to provide a comprehensive evaluation of genetic variation while being more cost-effective than piece-meal genetic testing. These technological advances can thus allow patients to end their diagnostic odysseys faster and cheaper than ever before.

Given the potential benefits of exome sequencing in providing diagnoses for short stature patients, we proposed a diagnostic flowchart for patients presenting with short stature diagnostic flowchart suggested that patients with short stature should undergo a limited panel of specific genetic tests for mutations consistent with their phenotype (e.g., microcephaly panel for a patient with microcephaly and short stature). We then suggested that patients who are negative for this set of specific tests should undergo comprehensive evaluation by exome sequencing for coding mutations and chromosomal microarrays for assessment of copy number variation. Whether our diagnostic workflow will result in higher diagnostic rates for patients with short stature remains to be determined.

18.5 Copy Number Variation

Thus far, this chapter has focused on simpler forms of genetic variation—SNPs and short indels. However, larger insertions and deletions referred to as copy number variation (CNV), or structural variations (SVs) in the genome such as inversions can also play an important role in disease (McCarroll and Altshuler 2007; Zhang et al. 2009). In fact, the number of base pairs in a person's genome that is perturbed by CNV/SVs is much greater than that perturbed by simple SNPs and indels (Handsaker et al. 2015).

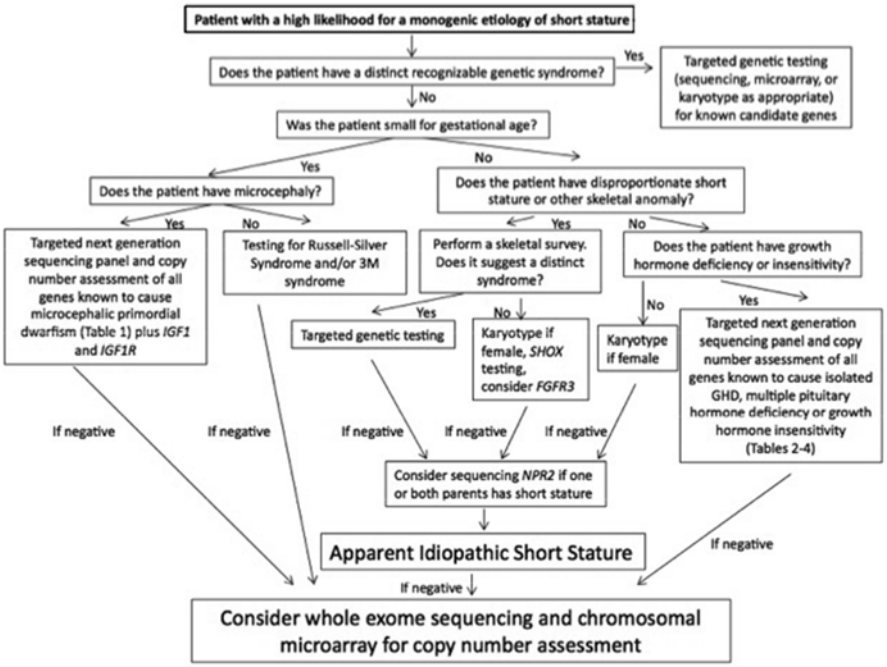


Fig. 18.1 Proposed diagnostic flowchart for patients being evaluated for short stature (Figure adapted from Dauber (an author of this chapter), Rosenfeld, and Hirschhorn, JCEM, 2014 (PMID 24915122))

CNVs and SVs have been studied for many decades in the field of genetics, as for a long time, they were the only forms of genetic variation that could be detected using available cytogenetic technologies. At the extremes of size, CNVs and SVs can be visible under a microscope using standard karyotyping methods. For many years, genomic lesions in patients were identified by cytogeneticists using karyotyping. However, these studies were incredibly laborious and had limited resolution (>10 million base pairs). In the late 1990s, a new high throughput method called array comparative genomic hybridization (array CGH) was developed, which allowed for highly accurate and cheap detection of copy number changes (Pinkel et al. 1998). The technology also allowed for much greater resolution, down to approximately 100 kilobases. Array CGH uses probes that bind throughout the genome, and compares the binding intensities at each probe compared to diploid control signals. More recently, technologies have been developed to identify copy number variation from exome or whole genome sequencing data (Zhao et al. 2013). These approaches generally utilize read depth intensities or split reads (reads mapping to discordant places in the genome) to identify CNVs or SVs.

While there are clear examples of CNVs that cause Mendelian growth disorders, we sought to test whether CNVs might contribute to short stature in a general population. Using a cohort of 2411 children who had array CGH data, we found that individuals with short stature had a greater burden of rare (<1 %) or lower-frequency (<5 %) deletions in their genome (Dauber et al. 2011). The average CNV length was

also higher in the individuals with short stature. Interestingly, there was no association between the presence of duplications and stature, nor was there an association between CNVs and tall stature. These results were replicated in a population-based cohort of nearly 7000 individuals. A regional CNV association study was also performed and identified three common CNVs that were associated with stature. In each of these regions, deletions were associated with decreased stature, while duplications were associated with increased stature. It is not as of yet clear how these deletions result in short stature, though it is likely through decreased dosage of important genes for growth and/or unmasking of recessive alleles (Joshi et al. 2015).

While overall burden of deletions might be associated with short stature, there is also value to identifying individual CNVs that cause growth disorders. In a patient with paternally-inherited short stature and microcephaly, array CGH performed on the patient and his parents revealed a heterozygous deletion of *IGF1* that was inherited from the father (Batey et al. 2014). Mutations and deletions in *IGF1* had previously been linked to short stature and microcephaly. The molecular diagnosis of *IGF1* deletion was consistent with the patient's phenotype. This patient was also the first to be reported with a full *IGF1* deletion, helping to clarify the phenotypic effects of haploinsufficiency of *IGF1*. Additionally, this patient demonstrates the utility of assaying CNVs clinically in the diagnosis of growth disorders.

18.6 SNP Genotyping and Genome Wide Association Studies

While the majority of this chapter has been focused on studies of rare variants with large phenotypic effect, there is also tremendous value to studying more common variants (those present in many people in the population) (Hirschhorn 2009; Visscher et al. 2012). Common variants typically have much smaller phenotypic effect, since if they had large effects on disease, they would be rapidly eliminated by natural selection (Manolio et al. 2009). Nonetheless, common variants can be incredibly important to disease risk and other human traits. Studies have suggested that the majority of variation in height can be explained by common genetic variation (Yang et al. 2015). Thus, while the contribution of each common genetic variant may only have a small effect on height, collectively they amount to a sizeable contribution, since each person will carry many common variants with effects that add up (Wood et al. 2014). Moreover, the genes identified by these common variants can have very important roles in growth biology, despite the common variants themselves each having small effects.

Common variants can be assessed rapidly and cheaply using SNP genotyping chips. These genotyping chips generally assay 100,000 to 1 million pre-specified common genetic variants across the genome, at a fraction of the cost of a whole genome sequence. Using haplotype maps, which are catalogs of the patterns of genetic variation across populations, one can then fairly accurately impute (infer) the state of all common variants (~10 million) across the genome (1000 Genomes Project Consortium et al. 2015; International HapMap 3 Consortium et al. 2010). Moreover, recent developments have allowed for imputation of even rarer variants (down to 0.1 % allele frequency). These SNP genotypes, whether directly geno-

typed or imputed, can then be tested for their association with a phenotype such as height in large populations.

GWAS for height are among the largest conducted to date. The most recent effort led by the GIANT consortium analyzed 250,000 individuals for the association between SNP genotypes and height. The analysis identified 697 variants independently associated with height; these variants were distributed across 423 loci or regions of the genome (Wood et al. 2014). The GWAS identified many known genes and pathways for growth, including fibroblast growth factor signaling and cartilage development. Importantly, many of these association signals did not point to genes

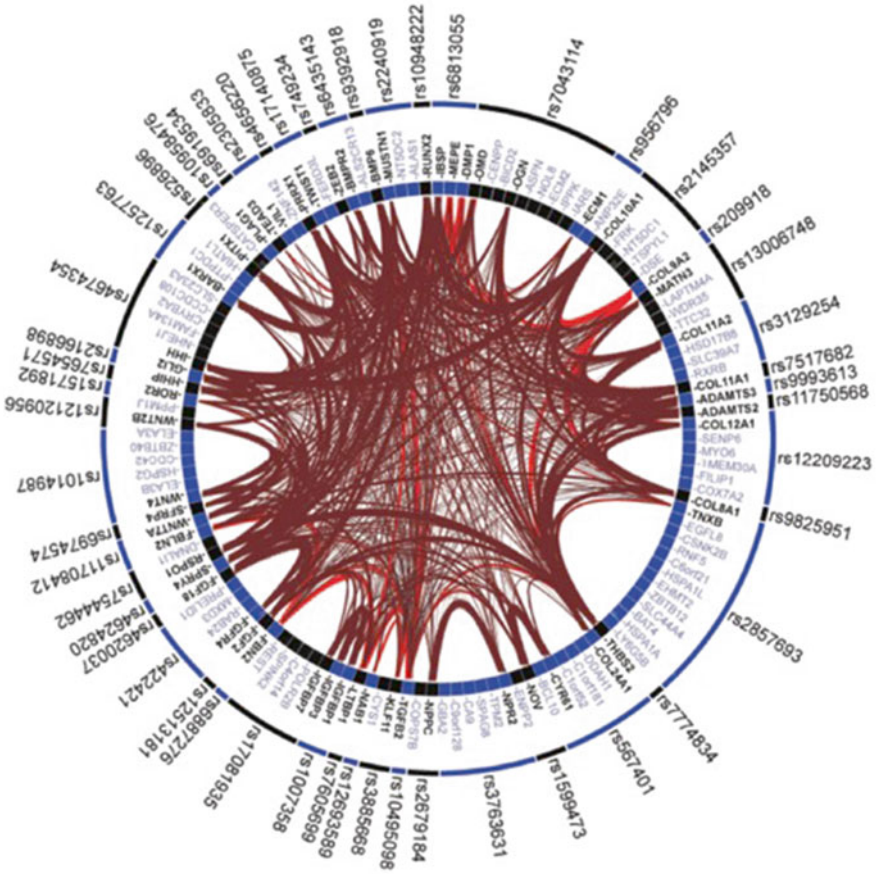


Fig. 18.2 Literature mining analysis of genes implicated by GWAS of height. A subset of the 697 height-associated SNPs are arranged along the *outer circle*. The *inner circle* represents the individual prioritized genes at leach locus implicated by these SNPs. Gene names shown in *black* have literature connections, and the ones in *gray* do not. The *redness* and thickness of the lines connecting pairs of genes represent the strength of the connections. Top-ranking keywords prioritized by GRAIL were 'transcription', 'growth', 'nuclear', 'factor', 'binding', 'collagen', 'differentiation', 'promoter', 'development', 'ribosomal', 'bone', 'mice', 'expression' and 'cartilage' (Figure from Wood et al., *Nature Genetics,* 2014 (PMID: 25282103))

with known function in human growth, generating many novel biological hypotheses about new genes involved in growth. Thus, GWAS offers a treasure trove of biological hypotheses that can be followed-up on with further genetic or functional studies.

As common variants have such an important influence on height, it has been suggested that these GWAS results can be used to predict height (or risk of diseases such as type 2 diabetes). By knowing the effect of each genetic variant on height and by measuring the state of each variant in an individual, one can then obtain an estimate of height based on these SNPs. However, to date, there appears to be little utility to using GWAS results to predict height or other traits or diseases (Kraft and Hunter 2009). This is because for reasons that are as of yet unknown, GWAS have only identified a small proportion of the contribution of common genetic variants to height and other diseases and traits (Zuk et al. 2012). For height, despite remarkable sample sizes and 697 associations, only about one fifth of the variation in height has been explained (Wood et al. 2014). Thus, the GWAS results offer limited predictive value and are far inferior to family history or parental height. In the future however, GWAS results may explain a much greater fraction of the contribution of genetic variation to human traits and diseases and thus be a much better source for risk prediction.

18.7 Future Directions

While next generation genomic technologies have contributed to a rapid expansion in our knowledge about growth disorders and our ability to diagnose them, significant work remains to be done. Here, we outline some unsolved problems and applications of genomic technologies that are underway and are likely to play an important role in continuing to help unravel, and ultimately treat, growth disorders over the coming years.

While the genetic basis of 400 Mendelian growth disorders has been identified, there are many growth disorders where the genetic basis is as of yet unknown. Comprehensive analysis of the genetics of these patients using technologies such as next generation sequencing and array CGH will likely uncover many new causes of growth disorders. As the cost of these technologies falls rapidly, the capacity to analyze many more samples will increase. With a greater knowledge of the genes that can cause disease, our ability to diagnose patients with growth disorders will also likely improve.

For the most part, the identification of novel disease genes for Mendelian disorders has focused on the fortuitous observation of multiple patients with consistent phenotypes, who all carry mutations in a given gene. However, these observations often lack statistical and methodological rigor. Current and future research will likely increasingly apply analyses across cohorts of patients with a given phenotype and perform large scale sequencing analyses to identify statistically significant enrichments of rare potentially pathogenic variants in genes as compared to suitable

controls (Lee et al. 2014b). These approaches have been applied successfully to many other disorders, such as amyotrophic lateral sclerosis (ALS) (Cirulli et al. 2015). However, these methods, while statistically rigorous, will often require large sample sizes, especially when there are many genes that can cause a disease (high locus heterogeneity).

Although many genes are known for growth disorders and many more are likely to be discovered in the coming years, interpreting genetic variants in these genes remains difficult. Many patients will be sequenced and variants of uncertain significance will be found in these genes. In order for genomic technologies to be more useful in a clinical setting, we will need to greatly improve our ability to interpret genetic variants and distinguish benign variants from pathogenic. Computational algorithms will likely improve, especially for the noncoding portions of the genome (Ionita-Laza et al. 2016; Kircher et al. 2014). In addition, high throughput functional assays to test the effect of many different mutations might be generated for each gene, allowing us to determine *a priori* the functional consequence of every mutation in a given gene. This paradigm has already been applied for *PPARG* for type 2 diabetes and *BRCA1* for breast cancer risk (Majithia et al. 2014; Starita et al. 2015).

Genetics also has tremendous potential to help us understand biology and generate therapeutic hypotheses. However, for this to happen, we will need to translate genetic findings into the laboratory to understand how these genes and mutations cause growth disorders. These genes will likely become prime therapeutic targets, and genetics can offer a unique window into identifying therapeutic targets that are likely to be effective and safe (Plenge et al. 2013). This is because patients carrying genetic mutations approximate the effect of a therapeutic knocking down the effect of a given gene. The story of *PCSK9* demonstrates this paradigm. Individuals with mutations resulting in loss of function of *PCSK9* have very low LDL cholesterol levels and greatly decreased risk of heart attacks and are otherwise healthy (Cohen et al. 2005, 2006; Kotowski et al. 2006). Thus, targeting of PCSK9 with a therapeutic agent should result in decreased LDL cholesterol and heart attack risk with few side effects. Multiple clinical trials for monoclonal antibodies against PCSK9 have in fact been demonstrated to be effective and safe (Blom et al. 2014; Robinson et al. 2015; Sabatine et al. 2015). It is the hope of human genetics that this paradigm might be applied for many other human diseases.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526:68–74.

Ahmed ML, Allen AD, Sharma A, Macfarlane JA, Dunger DB. Evaluation of a district growth screening programme: the Oxford growth study. Arch Dis Child. 1993;69:361–5.

Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008;322:881–8.

Aspberg A. The different roles of aggrecan interaction domains. J Histochem Cytochem. 2012;60:987–96.

Baron J, Savendahl L, De Luca F, Dauber A, Phillip M, Wit JM, Nilsson O. Short and tall stature: a new paradigm emerges. Nat Rev Endocrinol. 2015;11:735–46.

Bartels CF, Bukulmez H, Padayatti P, Rhee DK, van Ravenswaaij-Arts C, Pauli RM, Mundlos S, Chitayat D, Shih LY, Al-Gazali LI, et al. Mutations in the transmembrane natriuretic peptide receptor NPR-B impair skeletal growth and cause acromesomelic dysplasia, type Maroteaux. Am J Hum Genet. 2004;75:27–34.

Batey L, Moon JE, Yu Y, Wu B, Hirschhorn JN, Shen Y, Dauber A. A novel deletion of IGF1 in a patient with idiopathic short stature provides insight Into IGF1 haploinsufficiency. J Clin Endocrinol Metab. 2014;99:E153–9.

Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova JL, Abel L. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A. 2015;112:5473–8.

Blom DJ, Hala T, Bolognese M, Lillestol MJ, Toth PD, Burgess L, Ceska R, Roth E, Koren MJ, Ballantyne CM, et al. A 52-week placebo-controlled trial of evolocumab in hyperlipidemia. N Engl J Med. 2014;370:1809–19.

Chan Y, Holmen OL, Dauber A, Vatten L, Havulinna AS, Skorpen F, Kvaloy K, Silander K, Nguyen TT, Willer C, et al. Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals. PLoS Genet. 2011;7:e1002439.

Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T, et al. The genetic basis of mendelian phenotypes: discoveries, challenges, and opportunities. Am J Hum Genet. 2015;97:199–215.

Cirulli ET, Lasseigne BN, Petrovski S, Sapp PC, Dion PA, Leblond CS, Couthouis J, Lu YF, Wang Q, Krueger BJ, et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. Science. 2015;347:1436–41.

Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet. 2005;37:161–5.

Cohen JC, Boerwinkle E, Mosley Jr TH, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 2006;354:1264–72.

Dauber A, Yu Y, Turchin MC, Chiang CW, Meng YA, Demerath EW, Patel SR, Rich SS, Rotter JI, Schreiner PJ, et al. Genome-wide association of copy-number variation reveals an association between short stature and the presence of low-frequency genomic deletions. Am J Hum Genet. 2011;89:751–9.

Dauber A, Golzio C, Guenot C, Jodelka FM, Kibaek M, Kjaergaard S, Leheup B, Martinet D, Nowaczyk MJ, Rosenfeld JA, et al. SCRIB and PUF60 are primary drivers of the multisystemic phenotypes of the 8q24.3 copy-number variant. Am J Hum Genet. 2013a;93:798–811.

Dauber A, Stoler J, Hechter E, Safer J, Hirschhorn JN. Whole exome sequencing reveals a novel mutation in CUL7 in a patient with an undiagnosed growth disorder. J Pediatr. 2013b;162:202–4.e1.

Dauber A, Ercan A, Lee J, James P, Jacobs PP, Ashline DJ, Wang SR, Miller T, Hirschhorn JN, Nigrovic PA, Sackstein R. Congenital disorder of fucosylation type 2c (LADII) presenting with short stature and developmental delay with minimal adhesion defect. Hum Mol Genet. 2014a;23:2880–7.

Dauber A, Rosenfeld RG, Hirschhorn JN. Genetic evaluation of short stature. J Clin Endocrinol Metab. 2014b;99:3080–92.

Dauber A, Munoz-Calvo MT, Barrios V, Domene HM, Kloverpris S, Serra-Juhe C, Desikan V, Pozo J, Muzumdar R, Martos-Moreno GA, et al. Mutations in pregnancy-associated plasma protein A2 cause short stature due to low IGF-I availability. EMBO Mol Med. 2016;8:363–74.

David A, Hwa V, Metherell LA, Netchine I, Camacho-Hubner C, Clark AJ, Rosenfeld RG, Savage MO. Evidence for a continuum of genetic, phenotypic, and biochemical abnormalities in children with growth hormone insensitivity. Endocr Rev. 2011;32:472–97.

de Bruin C, Dauber A. Genomic insights into growth and its disorders: an update. Curr Opin Endocrinol Diabetes Obes. 2016;23:51–6.

de Bruin C, Mericq V, Andrew SF, van Duyvenvoorde HA, Verkaik NS, Losekoot M, Porollo A, Garcia H, Kuang Y, Hanson D, et al. An XRCC4 splice mutation associated with severe short stature, gonadal failure, and early-onset metabolic syndrome. J Clin Endocrinol Metab. 2015;100:E789–98.

de Villartay JP. When natural mutants do not fit our expectations: the intriguing case of patients with XRCC4 mutations revealed by whole-exome sequencing. EMBO Mol Med. 2015;7:862–4.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.

Etzioni A, Frydman M, Pollack S, Avidor I, Phillips ML, Paulson JC, Gershoni-Baruch R. Brief report: recurrent severe infections caused by a novel leukocyte adhesion deficiency. N Engl J Med. 1992;327:1789–92.

Faiyaz-Ul-Haque M, Zaidi SH, Al-Ali M, Al-Mureikhi MS, Kennedy S, Al-Thani G, Tsui LC, Teebi AS. A novel missense mutation in the galactosyltransferase-I (B4GALT7) gene in a family exhibiting facioskeletal anomalies and Ehlers-Danlos syndrome resembling the progeroid type. Am J Med Genet A. 2004;128A:39–45.

Gleghorn L, Ramesar R, Beighton P, Wallis G. A mutation in the variable repeat region of the aggrecan gene (AGC1) causes a form of spondyloepiphyseal dysplasia associated with severe, premature osteoarthritis. Am J Hum Genet. 2005;77:484–90.

Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics. 2013;14 (Suppl 3):S7-2164-14-S3-S7. Epub 28 May 2013.

Golan D, Erlich Y, Rosset S. Weighted pooling—practical and cost-effective techniques for pooled high-throughput sequencing. Bioinformatics. 2012;28:i197–206.

Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, Sunyaev S. Sequencing studies in human genetics: design and interpretation. Nat Rev Genet. 2013;14:460–70.

Green AA, MacFarlane JA. Method for the earlier recognition of abnormal stature. Arch Dis Child. 1983;58:535–7.

Grimberg A, Kutikov JK, Cucchiara AJ. Sex differences in patients referred for evaluation of poor growth. J Pediatr. 2005;146:212–16.

Grote FK, Oostdijk W, De Muinck Keizer-Schrama SM, van Dommelen P, van Buuren S, Dekker FW, Ketel AG, Moll HA, Wit JM. The diagnostic work up of growth failure in secondary health care; an evaluation of consensus guidelines. BMC Pediatr. 2008;8:21-2431-8-21.

Guo MH, Stoler J, Lui J, Nilsson O, Bianchi DW, Hirschhorn JN, Dauber A. Redefining the progeroid form of Ehlers-Danlos syndrome: report of the fourth patient with B4GALT7 deficiency and review of the literature. Am J Med Genet A. 2013;161A:2519–27.

Guo MH, Shen Y, Walvoord EC, Miller TC, Moon JE, Hirschhorn JN, Dauber A. Whole exome sequencing to identify genetic causes of short stature. Horm Res Paediatr. 2014;82:44–52.

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA. Large multiallelic copy number variations in humans. Nat Genet. 2015;47:296–303.

Hanson D, Murray PG, Black GC, Clayton PE. The genetics of 3-M syndrome: unravelling a potential new regulatory growth pathway. Horm Res Paediatr. 2011;76:369–78.

Hastings ML, Allemand E, Duelli DM, Myers MP, Krainer AR. Control of pre-mRNA splicing by the general splicing factors PUF60 and U2AF(65). PLoS One. 2007;2:e538.

Hicks S, Wheeler DA, Plon SE, Kimmel M. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum Mutat. 2011;32:661–8.

Hirschhorn JN. Genomewide association studies—illuminating biologic pathways. N Engl J Med. 2009;360:1699–701.
Hirschhorn JN, Lettre G. Progress in genome-wide association studies of human height. Horm Res. 2009;71 Suppl 2:5–13.
International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467:52–8.
Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48:214–20.
Joshi PK, Esko T, Mattsson H, Eklund N, Gandin I, Nutile T, Jackson AU, Schurmann C, Smith AV, Zhang W, et al. Directional dominance on stature and cognition in diverse human populations. Nature. 2015;523:459–62.
Kannu P, Bateman J, Savarirayan R. Clinical phenotypes associated with type II collagen mutations. J Paediatr Child Health. 2012;48:E38–43.
Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–15.
Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005;308:385–9.
Klingseisen A, Jackson AP. Mechanisms and pathways of growth failure in primordial dwarfism. Genes Dev. 2011;25:2011–24.
Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155:27–38.
Kotowski IK, Pertsemlidis A, Luke A, Cooper RS, Vega GL, Cohen JC, Hobbs HH. A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. Am J Hum Genet. 2006;78:410–22.
Kraft P, Hunter DJ. Genetic risk prediction—are we there yet? N Engl J Med. 2009;360:1701–3.
Kresse H, Rosthoj S, Quentin E, Hollmann J, Glossl J, Okada S, Tonnesen T. Glycosaminoglycan-free small proteoglycan core protein is secreted by fibroblasts from a patient with a syndrome resembling progeroid. Am J Hum Genet. 1987;41:436–53.
Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, Das K, Toy T, Harry B, Yourshaw M, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA. 2014a;312:1880–7.
Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014b;95:5–23.
Lindsay R, Feldkamp M, Harris D, Robertson J, Rallison M. Utah Growth Study: growth standards and the prevalence of growth hormone deficiency. J Pediatr. 1994;125:29–35.
Liu X, Han S, Wang Z, Gelernter J, Yang BZ. Variant callers for next-generation sequencing data: a comparison study. PLoS One. 2013;8:e75619.
Majithia AR, Flannick J, Shahinian P, Guo M, Bray MA, Fontanillas P, Gabriel SB, GoT2D Consortium, NHGRI JHS/FHS Allelic Spectrum Project, SIGMA T2D Consortium, et al. Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. Proc Natl Acad Sci U S A. 2014;111:13127–32.
Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.
Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto, Calif). 2013;6:287–303.
Marquardt T, Brune T, Luhn K, Zimmer KP, Korner C, Fabritz L, van der Werft N, Vormoor J, Freeze HH, Louwen F, et al. Leukocyte adhesion deficiency II syndrome, a generalized defect in fucose metabolism. J Pediatr. 1999;134:681–8.
McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet. 2007;39:S37–42.

Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinf. 2014;15:247-2105-15-247.
Mielczarek M, Szyda J. Review of alignment and SNP calling algorithms for next-generation sequencing data. J Appl Genet. 2016;57:71–9.
Minikel EV, Vallabh SM, Lek M, Estrada K, Samocha KE, Sathirapongsasuti JF, McLean CY, Tung JY, Yu LP, Gambetti P, et al. Quantifying prion disease penetrance using large population control cohorts. Sci Transl Med. 2016;8:322ra9.
Murray JE, van der Burg M, IJspeert H, Carroll P, Wu Q, Ochi T, Leitch A, Miller ES, Kysela B, Jawad A, et al. Mutations in the NHEJ component XRCC4 cause primordial dwarfism. Am J Hum Genet. 2015;96:412–24.
Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015;47:856–60.
Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461:272–6.
Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12:443–51.
Nilsson O, Guo MH, Dunbar N, Popovic J, Flynn D, Jacobsen C, Lui JC, Hirschhorn JN, Baron J, Dauber A. Short stature, accelerated bone maturation, and early growth cessation due to heterozygous aggrecan mutations. J Clin Endocrinol Metab. 2014;99:E1510–18.
Olney RC, Bukulmez H, Bartels CF, Prickett TC, Espiner EA, Potter LR, Warman ML. Heterozygous mutations in natriuretic peptide receptor-B (NPR2) are associated with short stature. J Clin Endocrinol Metab. 2006;91:1229–32.
Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet. 1998;20:207–11.
Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nat Rev Drug Discov. 2013;12:581–94.
Pulst SM. Genetic linkage analysis. Arch Neurol. 1999;56:667–72.
Quintos JB, Guo MH, Dauber A. Idiopathic short stature due to novel heterozygous mutation of the aggrecan gene. J Pediatr Endocrinol Metab. 2015;28:927–32.
Robinson DG, Storey JD. subSeq: determining appropriate sequencing depth through efficient read subsampling. Bioinformatics. 2014;30:3424–6.
Robinson JG, Farnier M, Krempf M, Bergeron J, Luc G, Averna M, Stroes ES, Langslet G, Raal FJ, El Shahawy M, et al. Efficacy and safety of alirocumab in reducing lipids and cardiovascular events. N Engl J Med. 2015;372:1489–99.
Sabatine MS, Giugliano RP, Wiviott SD, Raal FJ, Blom DJ, Robinson J, Ballantyne CM, Somaratne R, Legg J, Wasserman SM, et al. Efficacy and safety of evolocumab in reducing lipids and cardiovascular events. N Engl J Med. 2015;372:1500–9.
Shaheen R, Faqeih E, Ansari S, Abdel-Salam G, Al-Hassnan ZN, Al-Shidi T, Alomar R, Sogaty S, Alkuraya FS. Genomic analysis of primordial dwarfism reveals novel disease genes. Genome Res. 2014;24:291–9.
Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, Dunkel L, De Lange M, Harris JR, Hjelmborg JV, et al. Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res. 2003;6:399–408.
Sisley S, Trujillo MV, Khoury J, Backeljauw P. Low incidence of pathology detection and high cost of screening in the evaluation of asymptomatic short children. J Pediatr. 2013;163:1045–51.
Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, Fields S. Massively parallel functional analysis of BRCA1 RING domain variants. Genetics. 2015;200:413–22.
Stattin EL, Wiklund F, Lindblom K, Onnerfjord P, Jonsson BA, Tegner Y, Sasaki T, Struglics A, Lohmander S, Dahl N, Heinegard D, Aspberg A. A missense mutation in the aggrecan C-type

lectin domain disrupts extracellular matrix interactions and causes dominant familial osteochondritis dissecans. Am J Hum Genet. 2010;86:126–37.
Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187:367–83.
Tompson SW, Merriman B, Funari VA, Fresquet M, Lachman RS, Rimoin DL, Nelson SF, Briggs MD, Cohn DH, Krakow D. A recessive skeletal dysplasia, SEMD aggrecan type, results from a missense mutation affecting the C-type lectin domain of aggrecan. Am J Hum Genet. 2009;84:72–9.
Unger S, Gorna MW, Le Bechec A, Do Vale-Pereira S, Bedeschi MF, Geiberger S, Grigelioniene G, Horemuzova E, Lalatta F, Lausch E, et al. FAM111A mutations result in hypoparathyroidism and impaired skeletal development. Am J Hum Genet. 2013;92:990–5.
Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24.
Voss LD, Mulligan J, Betts PR, Wilkin TJ. Poor growth in school entrants as an index of organic disease: the Wessex growth study. BMJ. 1992;305:1400–2.
Wang SR, Carmichael H, Andrew SF, Miller TC, Moon JE, Derr MA, Hwa V, Hirschhorn JN, Dauber A. Large-scale pooled next-generation sequencing of 1077 genes to identify genetic causes of short stature. J Clin Endocrinol Metab. 2013;98:E1428–37.
Wang SR, Jacobsen CM, Carmichael H, Edmund AB, Robinson JW, Olney RC, Miller TC, Moon JE, Mericq V, Potter LR, et al. Heterozygous mutations in natriuretic peptide receptor-B (NPR2) gene as a cause of short stature. Hum Mutat. 2015;36:474–81.
Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C, et al. Comprehensive variation discovery in single human genomes. Nat Genet. 2014;46:1350–5.
Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46:1173–86.
Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, Ward P, Braxton A, Wang M, Buhay C, et al. Molecular findings among patients referred for clinical whole-exome sequencing. JAMA. 2014;312:1870–9.
Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH, Robinson MR, Perry JR, Nolte IM, van Vliet-Ostaptchouk JV, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015;47:1114–20.
Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451–81.
Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinf. 2013;14(Suppl 11):S1-2105-14-S11-S1. Epub 13 Sep 2013.
Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci U S A. 2012;109:1193–8.

Chapter 19
Systems Biology Approaches for Elucidation of the Transcriptional Regulation of Pulmonary Maturation

Yan Xu and Jeffrey A. Whitsett

Abstract Surfactant deficiency associated with lung immaturity is a major cause of morbidity and mortality in preterm infants resulting in acute respiratory failure (termed "respiratory distress syndrome" or RDS) and chronic respiratory dysfunction (bronchopulmonary dysplasia or BPD). The lack of surfactant lipids and proteins needed to reduce surface tension at the air-liquid interface in the peripheral lung saccules, causes atelectasis and respiratory insufficiency after preterm birth. Lung "maturation" is a complex process involving diverse structural, cellular, and biochemical changes in lung architecture and function that are precisely coordinated by genetic and environmental factors that synchronize the length of gestation with lung maturation. We developed computational tools utilizing systems biology and single cell genomics strategies to analyze large-scale mRNA expression data from distinct technical platforms and biological contexts to: (1) identify signature genes, (2) predict transcriptional regulators driving epithelial cell differentiation, and (3) model transcriptional regulatory networks (TRN) controlling perinatal lung maturation. The signaling and transcriptional mechanisms controlling lung growth and maturation required for the abrupt adaptation to airbreathing at birth are of considerable clinical interest, with the hope of preventing and treating neonatal pulmonary disease. This chapter provides practical analytic strategies to utilize bioinformatics tools to integrate large-scale lung gene expression data with independent genomic information to predict transcriptional networks controlling lung maturation and surfactant homeostasis.

Y. Xu (✉)
Department of Pediatrics and Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Divisions of Pulmonary Biology and Biomedical Informatics, University of Cincinnati College of Medicine,
3333 Burnet Avenue, MLC7009, Cincinnati, OH 45229-3039, USA
e-mail: Yan.Xu@cchmc.org

J.A. Whitsett
Department of Pediatrics, Cincinnati Children's Hospital Medical Center Perinatal Institute, Division of Neonatology, Perinatal, and Pulmonary Biology, University of Cincinnati College of Medicine, Cincinnati, OH 45229-3039, USA
e-mail: Jeffrey.whitsett@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_19

Keywords Lung maturation • Gene expression • Transcriptional regulatory networks (TRN) • Respiratory distress syndrome • Single cell RNA-seq (scRNA-seq)

19.1 Introduction

The timing of lung maturation is precisely controlled by complex genetic and cellular programs. Preterm infants, born less than 36 weeks, are at risk for respiratory distress syndrome (RDS) at birth, most commonly the cause of perinatal mortality and morbidity in preterm infants. Immaturity of alveolar type 2 (AT2) cells causes surfactant deficiency and atelectasis of respiratory failure after birth (Dubin 1990; Grenache and Gronowski 2006). Since the discovery that RDS is caused by the lack of pulmonary surfactant (Avery and Mead 1959), the structure, function and clinical relevance of surfactant lipids and proteins have been extensively studied *in vivo* and *in vitro* (Weaver and Beck 1999; Weaver and Whitsett 1991; Whitsett 2006; Whitsett et al. 1995; Whitsett and Weaver 2002). Although advances in clinical care of preterm infants, including surfactant replacement and administration of antenatal glucocorticoids to induce lung maturation have improved perinatal survival at ever earlier gestations, the increasing survival of extremely preterm infants and the challenges of their postnatal care have been associated with chronic lung injury and remodeling, resulting in increased number of infants with bronchopulmonary dysplasia (BPD). Preterm birth rates (11–12 %) and associated pulmonary morbidities remain a major cause of infant mortality worldwide, affecting approximately 500,000 babies in North America alone (Goldenberg et al. 2008; Gravett et al. 2010). The molecular and cellular mechanisms controlling preterm birth and lung maturation remain enigmatic (Muglia and Katz 2010). Progress in prevention and therapy of pulmonary immaturity associated with preterm birth will depend upon deeper understanding of the cellular, genetic, environmental, and hormonal factors controlling perinatal lung function.

Perinatal lung "maturation," necessary for the transition from intrauterine to extra-uterine life, is dependent upon the integration of structural, biochemical and physiologic factors affecting the differentiation and function of alveolar epithelial cells. AT2 cell differentiation determines the production of pulmonary surfactant, a complex mixture of lipids and proteins that is essential for reducing surface tension created at the air-liquid interface in the alveoli (Burri 1984; McMurtry 2002). Recent technical advances in RNA sequencing, high resolution imaging, and computational sciences afford the opportunity to extend present knowledge regarding the diversity of cells and the genetic programs that determine alveolar structure and function in the perinatal period. The availability of complete DNA sequences of the human, mouse and other genomes and the introduction of high throughput "Omics" technologies for expression profiling enable simultaneous quantification of RNAs, genes, metabolites, and proteins. Recent advances in single-cell isolation and massive parallel DNA sequencing enable resolution of gene expression in individual

cells, providing insight into the diversity of cell types, and the genetic networks directing cell differentiation, and the interactions among diverse cell types. These technologies bring new perspectives for the study of gene networks and their regulation at individual cell level, providing access to molecular mechanisms underlying various diseases and phenotypes. While phenotypic outcomes of experiments designed to regulate expression or function of genes of interest in animal models of disease have been useful for study of disease pathogenesis, systems biology, using computational modeling of multi-dimensional data to predict biological mechanisms (Bunyavanich and Schadt 2015; Kitano 2002), is providing increasing insight into the pathogenesis of human disease, including those affecting the lung. In this chapter, we discuss relatively new technologies and approaches being applied to the study of perinatal lung maturation, with emphasis on single cell transcriptomics and high-resolution imaging to understand the biological and morphological processes mediating lung development and maturation. We provide examples in which systems biology approaches have been used to integrate complex data from distinct technical platforms; our ability to integrate and interpret increasingly complex data will be key to improving our understanding of how cell behaviors are dynamically regulated to maintain normal lung structure and function.

19.2 Structural and Morphological Changes Associated with Lung Development

Lung development has been divided into five morphologically distinct stages that begin with formation of the primordial lung buds and continue through the postnatal period. In the mouse, the embryonic stage is distinguished by the formation of the trachea, lung buds and division of the trachea and esophagus (E9-11.5). Branching morphogenesis and vasculogenesis forming the conducting airways and peripheral acinar tubules and buds is highly active during the pseudoglandular stage (E11.5-15.5). During the canalicular stage (E15.5-17.5), increasing numbers of acinar tubules are formed, angiogenesis and vasculogenesis are active, and differentiation of alveolar type I and II epithelial cells (AT1 and AT2 cells) begins. In the saccular stage (E17.5- PN5), terminal respiratory saccules dilate, and mesenchymal components of the peripheral lung thin. During the alveolar stage (PN5-30), the alveolar-capillary network matures, and the processes of alveolar septation and alveolar growth proceed to produce the mature lung. (Maeda et al. 2007; Perl and Whitsett 1999). Lung maturation is a continuous processes involving proliferation, migration, and differentiation of a great diversity of cell types of neuronal, epithelial, mesenchymal and hematopoietic origins. Historically, the maturation of AT2 cells and associated production of pulmonary surfactant lipids and proteins has been the primary focus of studies to prevent and treat surfactant deficiency.

19.2.1 Structural Analysis of Lung Maturation at Tissue and Cellular Levels

Electron microscopy has been a powerful approach to identify the detailed architecture and accompanying lung development (Burri 1984; Ten Have-Opbroek 1991; Weibel 2015). Recent advances in high resolution confocal microscopy, immunochemistry, and new procedures to clarify tissue prior to fluorescence imaging, are enabling identification of specific pulmonary cell types and their precise anatomic positions during lung maturation (Amos and White 2003; Inerot et al. 1991; Kherlopian et al. 2008; St Croix et al. 2005; Vielreicher et al. 2013).

Figure 19.1 summarizes ontogenetic changes in lung architecture and epithelial differentiation during prenatal lung maturation. Immunofluorescence confocal microscopy was used to image fetal mouse lungs, spanning the canalicular saccular transition prior to birth to the postnatal alveolar stage. As shown in Fig. 19.1, During the canalicular to saccular transition at E16.5, proximal regions of the peripheral lung tubules are increasingly dilated. Epithelial cells in "transitional ducts" become

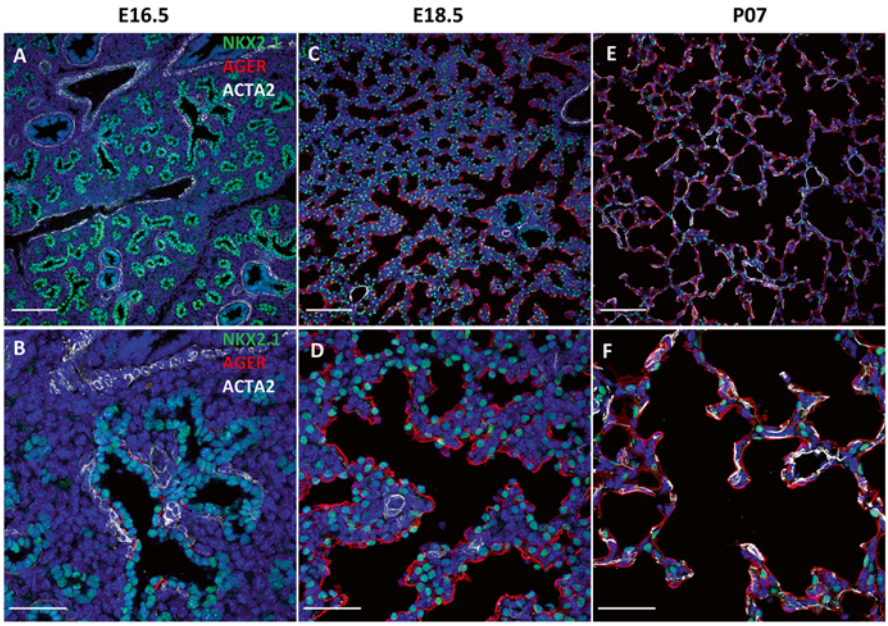


Fig. 19.1 Dynamic changes in tissue architecture associated with perinatal pulmonary maturation. Immunofluorescence confocal microscopy images of mouse lung sections from E16.5, 18.5, and postnatal day 7 are shown. Tissues were stained for NKX2-1, a marker of respiratory epithelial cells, ACTA2 (smooth muscle actin), and AGER, a marker of Alveolar Type 1 cell (AT1) cells. Dramatic changes in lung morphology occurs during perinatal lung maturation from the canalicular (E16.5), saccular (E18.5), and alveolar (PND7) stages of development. Note the thinning of mesenchymal tissue, dilation of terminal saccules, and septation of alveolar spaces seen in the postnatal tissue. Images are courtesy of Joseph Kitzmiller, CCHMC

more squamous and express Ager, an AT1 cell selection marker after birth. Cuboidal progenitor cells located in peripheral acinar buds at the ends of the lung tubules express surfactant proteins that are selective markers of AT2 cells. Mesenchymal components of the lung are relatively more abundant at early gestation and consist of multiple cell types. α-SMA, a marker of smooth muscle cells surrounding conducting airways (bronchial, bronchiolar), and vascular structures, and in lesser abundance, in the walls of peripheral lung saccules is dynamically regulated during alveolarization. During the saccular period (E18.5), peripheral lung saccules are increasingly dilated, and the proportion of mesenchyme to epithelium is decreased. After birth, septation of peripheral saccules occurs to create the thin alveolar structures characteristic of the mature lung. AT2 cells are readily distinguished from AT1 cells expressing AGER after birth. The utilization of cell-specific antisera, fluorescence probes, and transgenic mice that are useful for genetic labeling of cells for cell lineage analysis, are providing detailed information regarding the cellular processes mediating branching morphogenesis and alveolarization (Metzger et al. 2008).

19.3 Gene Expression Profiling Applications in Lung Development

The "transcriptome" is the complete complement of all RNA molecules produced in one or a population of cells. Qualitative, quantitative, and temporal analyses of transcriptomes provide opportunity to systematically understanding of organ development and disease. The invention of the DNA and RNA microarray chip technology enabled simultaneous identification and quantification of RNAs (DeFelice et al. 2003; Velculescu et al. 1997; Wang et al. 2000; Zuo et al. 2002). DNA-Sequencing has further enhanced genome-wide gene expression profiling and other high throughput 'omics' studies are enabling even deeper insights into specific biological processes and regulatory relationships among genes and cells (Bunyavanich and Schadt 2015; Burgess 2015; Kelly et al. 2013; Xu et al. 2010).

RNA microarray technology has been widely applied to many aspects of pulmonary biology (Mayburd et al. 2006; Minn et al. 2005; Wang et al. 2000; Zuo et al. 2002). Our own studies have contributed to the contemporary body of knowledge applying functional genomics approaches to study the transcriptional regulatory programs controlling lung development, function, and disease. Distinct sets of signaling molecules and transcription factors interact to implement the structural maturation and cell type specific differentiation of the lung. Through genetic manipulation of key signaling molecules and transcription factors in transgenic mouse models and the application of genome-wide transcriptional profiling analysis to these models, we have identified target genes, pathways, and physiologic consequences in response to the deletion or mutation of transcription factors such as NKX2-1, FOXA1/A2/A3, C/EBPα, HIF1α, STAT3, NFAT, SREBP, SPDEF, SOX17, PTEN,

NF1, KLF5, and FOXM1 and signaling molecules such as CNB1, MIA, SHH, CSF2R, FGF, and RSPO1 (Bell et al. 2011, 2013; Besnard et al. 2007, 2009; Bridges et al. 2014; Chen et al. 2009a, 2010, 2014; Dave et al. 2006; DeFelice et al. 2003; Lange et al. 2014; Lin et al. 2008; Maeda et al. 2011, 2012; Martis et al. 2006; Metzger et al. 2007; Miller et al. 2004; Rajavelu et al. 2015; Wan et al. 2004, 2005, 2008; Xu et al. 2003, 2006, 2007, 2009). Through those studies, new transcriptional target genes and networks were identified; provided insights into the molecular mechanisms underlying various lung diseases and phenotypes. Since key regulators of lung maturation are also critical for early embryogenesis, their disruption often results in embryonic lethality before lung formation. For the study of these genes, cell type specific, conditional mutagenesis has been useful in identifying the functions of transcription factors and receptor mediated signaling in lung morphogenesis, differentiation, and function. Multiple transcription factors and signaling pathways have been implicated in the structural and functional adaptation of the lung at birth (Maeda et al. 2007). Factors important for lung development and function are not exclusively lung-specific, tissue specificity being derived from the unique combinations and interactions of transcription factors. An understanding of the individual transcription factors and their interactions in the context of lung development requires systematic approaches to connect changes in gene expression with transcriptional networks governing lung maturation and differentiation (for review, see reference (Maeda et al. 2007)).

19.3.1 Next-Generation DNA Sequencing (NGS)

While RNA microarray analyses enabled qualitative and quantitative measures of gene expression, Next-generation DNA Sequencing (NGS) provides greater sensitivity and accuracy for transcriptional profiling (Kelly et al. 2013). NGS enables measurement of allele-specific gene expression, alternative start sites and splice variations, identification of single nucleotide polymorphisms, chromosomal translocations and fusions (Hitzemann et al. 2013). As costs of DNA sequencing decreased, and analytic methods matured, NGS is increasingly preferred for genomic applications.

19.3.2 Single Cell Transcriptomics

While analysis of high-throughput RNA expression profiles together with gene perturbation and targeting are providing an increasingly detailed insight regarding the genomic effects of important genes and regulators, the extent of cellular heterogeneity, and transitional states of cell differentiation and gene expression cannot be readily addressed by measuring transcripts derived from whole organs or pooled cell

populations. Ideally, addition of transcriptomic analyses at single cell resolution will provide further insights into processes of organogenesis and function.

Recent advances in microfluidics, robotics, amplification chemistries, and DNA sequencing technologies provide the ability to isolate, sequence, and quantitate RNA transcripts from single cells (Guo et al. 2010; Saliba et al. 2014; Shapiro et al. 2013; Yin and Marshall 2012). Single-cell transcriptomics is increasingly employed in the study of organ formation and function. For example, single-cell RNA-seq has been used to study differentiation of embryonic stem cells, neurons, lung epithelial cells and peripheral circulating tumor cells to identify novel lineage-specific markers and cell subtypes, and reveal dynamic changes in gene expression (Deng et al. 2014; Qiu et al. 2012; Ramskold et al. 2012; Tang et al. 2009, 2010; Trapnell et al. 2014; Treutlein et al. 2014). scRNA-seq has also been applied to isolated lung epithelial cells to predict epithelial lineages during development and after injury (Treutlein et al. 2014; Vaughan et al. 2015). Single-cell genomics is a powerful tool to profile cell-to-cell variability on a genomic scale. The transcriptome of genetically identical cells within a population may differ on the single-cell level; suggesting the dynamic and stochastic nature of gene expression "states" of individual cells (Saliba et al. 2014). Using single cells and/or sorted cells isolated from lung tissues at different development stages, we begin to map cellular processes underlying lung formation and maturation at single cell level.

19.4 Epigenetic Regulation at Single Cell Level

Epigenetic mechanisms, including histone modifications, DNA methylation, small and non-coding RNAs, and regulators of chromatin architecture play critical roles in regulating gene expression (Jaenisch and Young 2008; Vince et al. 2007). It is important to be able to profile the epigenome at single cell level, which then can be used to identify cell specific epigenetic biomarkers for clinical prediction, diagnosis, and therapeutic development. Recent breakthrough studies have shed the lights on this new emerging field, e.g., researchers at the Whitehead Institute have developed new methodology to monitor changes in DNA methylation over time in individual cells. The investigators developed a DNA methylation reporter system that mirrors whether a nearby region is methylated. When the target region is unmethylated, the reporter is also unmethylated, which allows expression and visualization of a fluorescent protein encoded by the reporter (Stelzer et al. 2015). Research by Bintu et al. measured effects of DNA methylation and histone modifications by methylation or deacetylation in single cells using time-lapse microscopy. Study showed that silencing was an all-or-none stochastic event. Furthermore, the duration of recruitment of the chromatin regulators determined the fraction of cells that were silenced. Thus, distinct modifiers can produce different characteristics of epigenetic memory (Bintu et al. 2016). Rapid advances in technologies to assess the impact of chromatin structure, the sites and regulation of enhancers and promoters via histone modifications, DNA methylation, hydroxymethylation and the role of

non-coding RNAs are transforming our understanding of gene regulation (Bintu et al. 2016; Jaenisch and Young 2008). With these newly developed methods, it is increasingly possible to map temporal and context dependent mechanisms regulating organogenesis.

19.5 Bioinformatics Approaches to Understand Perinatal Lung Maturation

The post-genomic era is providing high density genomic, proteomic and other high-throughput data relevant to lung biology. With thorough analysis, these data enable increasing insight into the complex biological processes underlying lung formation and function, but are challenged by the difficulty in causally linking changes in global gene expression to precise transcriptional mechanisms regulating cell behavior. Integrative approaches are needed to link gene expression, cell biology, and organ function. A wide range of statistical approaches is available for analysis of 'omic' data. Choices of analytic approaches depend both on experimental design and the nature of the 'omic' data. For reader's convenience, we listed some commonly used bioinformatics resources for functional genomic analyses and their corresponding Websites (Table 19.1). To decipher genomic responses of signaling molecules and transcription factors in gene perturbation studies, we developed integrative approaches to analyze large-scale mRNA expression data sets to identify transcriptional regulatory networks controlling dynamic processes of lung maturation and surfactant homeostasis (Besnard et al. 2011; Xu et al. 2010, 2012b).

19.5.1 Preparing Gene Lists for Comparisons

The most basic and common task in analyzing transcriptional profiling experiments is to identify differentially expressed genes influenced by experimental conditions. Fold change is a standard way to define differentially expressed genes (i.e., genes with ratios above a fixed cut-off, typically a difference of 1.5–2.0 being considered to be significant). Given the size of the experiments and extended data, false discovery rates are large in the study of RNA expression data that rely on fold change alone. Replication is an essential component of experimental design, enabling an estimation of variability and identification of biologically reproducible changes across samples. When using three or more replicates in each experimental condition, standard t-tests can be used to identify statistically significant changes between two groups. Multiple groups are compared via the ANOVA F statistic. For data lacking replicates or failing to satisfy normal distribution, model based methods such as DEseq, DEseq2 or EdgeR are preferred (Anders and Huber 2010; Love et al. 2014; Nikolayeva and Robinson 2014).

Table 19.1 Commonly used bioinformatics resources and their corresponding websites

Name	Source page	Functions	Category
Sincera	https://research.cchmc.org/pbge/sincera.html	*Single cell RNA-seq analysis to identify: cell types, cell type specific gene signatures; and driving forces of given cell types.*	*scRNA-seq*
LungGENS	https://research.cchmc.org/pbge/lunggens/default.html	*Single cell gene expression database, provide 'Query by single gene,' 'Query by gene list,' and 'Query by cell type' functions.*	*scRNA-seq*
BackSPIN	https://github.com/linnarsson-lab/BackSPIN	*scRNA-seq analysis: cell type identification (biclustering)*	*scRNA-seq*
Celloline / Cellity	https://github.com/Teichlab/celloline	*Mapping and quality assessment scRNA-seq data (Amazon Service required)*	*scRNA-seq*
ICGS	https://code.google.com/p/altanalyze/	*Correlated gene signatures analysis*	*scRNA-seq*
Monocle	http://cole-trapnell-lab.github.io/monocle-release/	*extracting lineage relationships from scRNA-seq and modeling the dynamic changes associated with cell differentiation*	*scRNA-seq*
RaceID	https://github.com/dgrun/RaceID	*scRNA-seq analysis: cell type identification (rare cell type)*	*scRNA-seq*
SCDE	http://hms-dbmi.github.io/scde/	*scRNA-seq analysis: differential expression analysis*	*scRNA-seq*
scLVM	https://github.com/PMBio/scLVM	*scRNA-seq analysis: clustering, confounding factor analysis*	*scRNA-seq*
SCUBA	https://github.com/gcyuan/SCUBA	*scRNA-seq analysis: cell ordering, lineage tree reconstruction, bifurcation analysis*	*scRNA-seq*
Singular	https://www.fluidigm.com/software	*scRNA-seq analysis: clustering, cell type identification, cell outlier detection*	*scRNA-seq*
SNN-Cliq	http://bioinfo.uncc.edu/SNNCliq/	*scRNA-seq analysis: clustering, cell type identification*	*scRNA-seq*

(continued)

Table 19.1 (continued)

Name	Source page	Functions	Category
Wanderlust	http://www.c2b2.columbia.edu/danapeerlab/html/wanderlust.html	*scRNA-seq analysis: cell ordering*	*scRNA-seq*
Waterfall	Supplemental of PMID 26299571	*scRNA-seq analysis: cell ordering, differential expression analysis (temporal)*	*scRNA-seq*
WikiPathway	http://www.wikipathways.org/index.php/WikiPathways	*Pathway search*	*Pathway*
JMP	http://www.jmp.com/en_us/home.html	*commercial statistical software*	*Other*
Orange	http://orange.biolab.si/	*Open source data visualization and data mining*	*Other*
AltAnalyze	http://www.altanalyze.org/	*Free Software Package (Gene expression and alternative splicing analyses on microarrays, RNASeq data and single cell data.)*	*NGS*
BamTools	https://github.com/pezmaster31/bamtools	*linux required.*	*NGS*
BRB-ArrayTools	http://brb.nci.nih.gov/BRB-ArrayTools/Documentation.html	*Free Software Package (Microarray gene expression, copy number, methylation and RNA-Seq data)*	*NGS*
DEGseq	https://www.bioconductor.org/packages/release/bioc/html/DEGseq.html	*R package, DE analysis*	*NGS*
DESeq	https://www.bioconductor.org/packages/release/bioc/html/DESeq.html	*R package, sequencing assays and DE analysis*	*NGS*
EdgeR	https://www.bioconductor.org/packages/release/bioc/html/edgeR.html	*R package, DE analysis*	*NGS*
FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/	*QA/QC*	*NGS*
Galaxy	https://usegalaxy.org/	*Free Software Package*	*NGS*
Genomatix	http://www.genomatix.de/index.html	*integrative software, commercial NGS data analysis, Genomic data mining*	*NGS*
NOISeq	https://www.bioconductor.org/packages/release/bioc/html/NOISeq.html	*R package, QA/QC,differential expression (DE) analysis*	*NGS*

Partek	http://www.partek.com/	*Commercial Software Suite (alignment, RNA-Seq, small RNA-Seq, DNA-Seq, and ChIP-Seq).*	*NGS*
Qualimap 2	http://qualimap.bioinfo.cipf.es/	*QA/QC, counts output*	*NGS*
rnaseqGene	http://www.bioconductor.org/help/workflows/rnaseqGene/	*R package: RNA-seq workflow, gene-level exploratory analysis and DE analysis*	*NGS*
SAMtools	https://sourceforge.net/projects/samtools/?source=navbar	*linux required.*	*NGS*
StrandNGS	http://www.strand-ngs.com/	*Commercial Software Suite (alignment, RNA-Seq, small RNA-Seq, DNA-Seq, Methyl-Seq, MeDIP-Seq, and ChIP-Seq)*	*NGS*
Cytoscape	http://www.cytoscape.org/	*Network Design and visulization*	*Graphics tool*
IGV	https://www.broadinstitute.org/igv/	*Genome viewer, splicing viewer*	*Graphics tool*
TreeView 3.0	http://jtreeview.sourceforge.net/	*Clustering, heatmap*	*Graphics tool*
Venny	http://bioinfogp.cnb.csic.es/tools/venny/	*Venn Digram, Overlapping lists.*	*Graphics tool*
Amigo2	http://amigo.geneontology.org/amigo	*Gene Set Enrichment Analysis*	*Enrichment Analysis*
DAVID	https://david.ncifcrf.gov/summary.jsp	*Gene Set Enrichment Analysis*	*Enrichment Analysis*
Onto-Express	http://vortex.cs.wayne.edu/Projects.html	*Gene Set Enrichment Analysis*	*Enrichment analysis*
Toppgene	https://toppgene.cchmc.org/	*Gene set enrichment analysis and candidate gene prioritization*	*Enrichment Analysis*
Allen Brain *Atlas*	http://www.brain-map.org/	*The Allen Institute for Brain Science*	*Database and resources*
COREMINE	http://www.coremine.com/medical/#search	*Comprehensive information on diseases, drugs, treatments and medical biology.*	*Database and resources*
EMBL-EBI	http://www.ebi.ac.uk/services	*Genome Browser, Knowledge Base*	*Database and resources*
ENCODE	http://genome.ucsc.edu/ENCODE/index.html	*Genome Browser, Knowledge Base*	*Database and resources*
Ensembl	http://www.ensembl.org/	*Genome Browser, Knowledge Base*	*Database and resources*
Gene Ontology Consortium	http://geneontology.org/	*Provide an up-to-date comprehensive gene ontology*	*Database and resources*

(continued)

Table 19.1 (continued)

Name	Source page	Functions	Category
GeneCards	http://www.genecards.org	*Genome Browser, Knowledge Base*	*Database and resources*
GenePaint	http://genepaint.org/Frameset.html	*GenePaint.org is a digital atlas of gene expression patterns in the mouse.*	*Database and resources*
UCSC Genome Browser	http://genome.ucsc.edu/cgi-bin/hgGateway	*Genome Browser, Knowledge Base*	*Database and resources*
GenomeNet	http://www.genome.jp/	*Genome Browser, Knowledge Base*	*Database and resources*
GUDMAP	http://www.gudmap.org/	*The GenitoUrinary Development Molecular Anatomy Project*	*Database and resources*
HGMD	http://www.hgmd.org	*gene lesions responsible for human inherited disease*	*Database and resources*
IPA	http://www.ingenuity.com/products/ipa	*Commercial Software Suite (Pathway analysis, literature mining)*	*Database and resources*
MGI	http://www.informatics.jax.org/	*Genome Browser, Knowledge Base*	*Database and resources*
NCBI	http://www.ncbi.nlm.nih.gov/	*Genome Browser, Knowledge Base*	*Database and resources*
OMIM	http://www.ncbi.nlm.nih.gov/omim	*human genes and genetic phenotypes*	*Database and resources*
The human protein atlas	http://www.proteinatlas.org/		*Database and resources*
UniProt	http://www.uniprot.org/uniprot/	*central hub for the of functional information on proteins*	*Database and resources*
Xenbase	http://www.xenbase.org/entry/	*Genome Browser, Knowledge Base*	*Database and resources*
ConsensusClusterPlus	https://www.bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html	*R package: clustering*	*Clustering*
GENE-E	http://www.broadinstitute.org/cancer/software/GENE-E/	*Clustering, heatmap*	*Clustering*
HCE	http://www.cs.umd.edu/hcil/hce/	*Clustering, heatmap*	*Clustering*
Tight clustering	http://rpackages.ianhowson.com/cran/tightClust/	*R package: clustering*	*Clustering*
BATS	http://bioinfo.na.iac.cnr.it/bats/index_file/download.htm	*Time course dataset analysis*	*Time Course*
STEM	http://www.sb.cs.cmu.edu/stem/	*Clustering of time course dataset*	*Time Course*

19.5.2 Pattern Recognition and Clustering Analysis

Pattern recognition is useful to group findings based on expression similarity and for summarizing and visualizing data. Dimension-reduction, e.g. principal components analysis (PCA), independent components analysis (ICA) and singular value decomposition (Holter et al. 2000) are commonly used. These methods are unsupervised, meaning that the information reduction is derived solely from the data and does not rely on previous knowledge or classifications.

Principal Components Analysis a commonly used dimension-reduction method, assumes that data variation can be explained by smaller numbers of transformed variables. PCA explains the variance-covariance structure of the original data through a few linear combinations of those variables and projects the data with thousands of dimensions into two- or three-dimensional spatial representations. While this is a highly useful method for data reduction, important data may be lost and in that case alternative approaches such as hierarchical clustering can be used.

Clustering Co-expression of multiple genes supports the likelihood that they share co-regulation by similar transcription factors or belong to same transcriptional network. Clustering provides insight into transcriptional networks by grouping genes on the basis of expression pattern that change similarly under various experimental conditions. Genes/proteins selected from "omics" data analyses can be grouped into distinct clusters. Genes in each cluster can be further classified according to Gene Ontology (GO) (http://geneontology.org/) and shared transcription factor binding sites (TFBS) in the regulatory regions of genes within the cluster to identify potential biological themes and common regulatory mechanisms represented by unique gene sets. Classical clustering algorithms including K-means, SOM and Hierarchical clustering are available in many commercial and open source analytic packages (e.g, Genespring, Partek, http://www.cs.umd.edu/hcil/hce/), more advanced algorithms recently developed, for example, "consensus clustering" and "tight clustering" enable more robust clustering results (Seo et al. 2006; Tseng and Wong 2005; Wilkerson and Hayes 2010). These methods generally emphasize clear group separations, and are assigned to only one cluster. In biology, genes, RNAs, proteins and metabolites may have multiple roles in cellular processes and respond to various conditions in complex ways. Fuzzy Heuristic Partition (Fu and Medico 2007; Gasch and Eisen 2002) considers each gene to be a member of every cluster with a variable degree of membership, enabling assignment of genes to more than one cluster with different degrees of membership. Using stringent membership cutoffs, genes in each cluster can be identified that are highly correlated across diverse experimental conditions. As the degree of membership decreases, additional genes join the cluster based on their expression similarity under distinct experimental conditions, enabling the identification of context-dependent regulation. We applied Fuzzy clustering algorithm (Fu and Medico 2007) to 194 mRNA microarray samples from 27 distinct mouse models and identified three closely related clusters were highly enriched in genes influencing lipid synthesis, lipid transport and surfactant homeostasis (Xu

et al. 2010). All three gene clusters share commonly enriched functions, i.e. "lipid biosynthesis and metabolism" and common transcription regulation by SREBP (*Srebf1* was present in all three gene clusters and was predicted to common upstream regulator of the three clusters). In addition to the shared function and regulation, each cluster has its uniquely enriched functionality. Cluster 1 is functionally enriched in "lung" and "vascular" development. Cluster 2 is enriched for "lipid metabolism and lipid transport"; "Endoplasmic reticulum (ER)" is the most enriched cellular component and genes in this cluster is most abundantly expressed in the lung. These functional annotations aligned well with the fact that surfactant lipid and proteins are synthesized and assembled in the ER of alveolar type II cells. Cluster 3 is featured by coupling "lipid metabolism" with "response to external/chemical stimulus", NFAT and STAT6 were predicted as unique regulators to C28 genes. Overall, the functional classifications indicate that lung lipid metabolism is closely associated with lung development and is required for various stress responses (Xu et al. 2010).

Other Usages of "Clustering" Clustering methods are most commonly used to group co-expressed genes; alternatively, to group genes based on their shared functional annotations, promoter/regulatory cis-elements, biochemical, and morphological features. Disease related subgroups, sharing phenotypes, genotypes or expression patterns, also can be used to cluster genes and processes involved in disease pathogenesis.

Figure 19.2 applies multivariate correlation analysis of mRNA expression data with lung physiology, biochemistry, and morphology. As depicted in Fig. 19.2a, qPCR data analysis of 53 mRNAs previously associated with lung function and structure formed three distinct clusters. Expression of cluster 1 genes, including surfactant and related proteins (*Sftpc, Abca3, and Slc34a2*), increased dramatically before birth; cluster 2 genes (*Nkx2-1, Pdgfa, Lpcat,* etc.) were moderately increased, while cluster 3 genes did not significantly change during the perinatal period. Using embryonic day 15 (E15) mRNA levels as baseline and E19.5 (the day before birth) as peak "maturation" patterns of RNA expression prior to birth were assessed. RNAs in cluster 1 & 2 were induced earlier and faster in B6 mice (born after 19.5 days gestation) than in A/J mice (born after 20.5 days gestation), indicating the dynamic expression changes in cluster 1 are closely associated with "shortened" gestation and earlier lung maturation. Multivariate correlation of mRNA expression data with dynamic changes of body weight, lung weight, saturated phosphatidylcholine (SatPC), an important component of surfactant, and morphometrics, identified a subset of mRNAs, including *Sftpa*, *Sftpb*, *Sftpc*, *Sftpd*, *Slc34a2*, *Scgb1a1*, *Cebpa* and *Aqp5*, that were highly correlated with these biochemistry and physiological indices of lung "maturation." Likewise, mRNAs associated with lipid homeostasis, including *Scd1*, *Abca3*, *Fabp5*, and *Lpcat1*, were correlated with lung weight and fractional area of airspace. In contrast, a distinct subset of mRNAs, including *Tubb3*, *Pygb*, and *Igfbp2*, were best correlated with the fractional area of

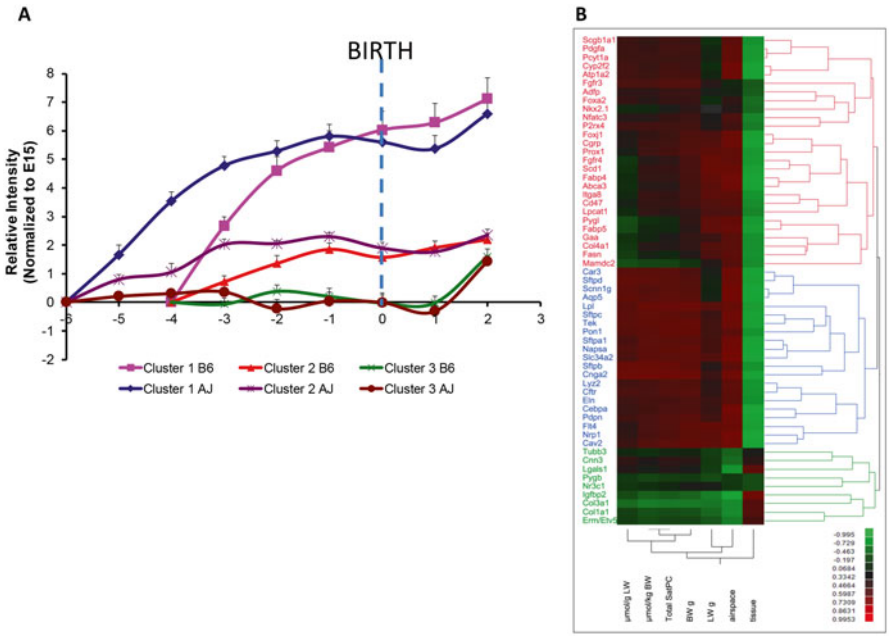
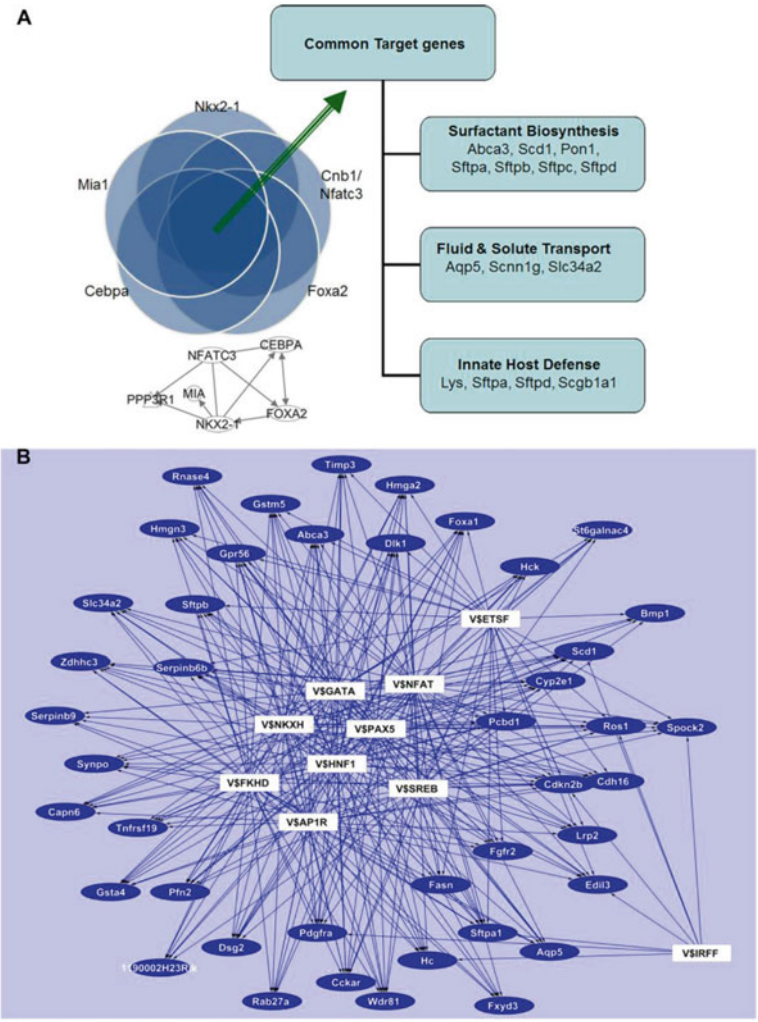


Fig. 19.2 Clustering selective lung maturation markers. (**a**) Dynamic changes in mRNA expression of 53 lung maturation markers were measured by qPCR in both A/J and B6 mouse strains. qPCR data analysis revealed three major clusters. The chart is generated based on the normalized cluster mean and standard error of each gene cluster. The X-axis represents days before and after birth (E15-P2): 0 is the day of birth. The Y-axis represents relative expression normalized to E15. (**b**) Dynamic mRNA expression profiles of 53 genes at different gestation ages for the A/J mice were correlated with body weight (BW), lung weight (LW), SatPC (μmol/gLW, μmol/gBW, and total), and morphometric measurements (airspace, tissue) at corresponding gestational ages using multivariate correction function from JMP 9 (SAS Institute Inc, NC). The heat map was generated based on data from A/J mice using Ward's minimum variance method to estimate cluster similarity. Gradients in the *red* and *green* color range indicate positive and negative correlation, respectively. The levels of mRNAs in red clusters were highly correlated with ontogenic changes in lung SatPC and fractional area of airspace mRNAs in *blue* clusters were moderately correlated with SatPC, but closely correlated with fractional area of airspace mRNAs in *green* clusters correlated well with the fractional area of the tissue compartment (Adapted from Figure S5 in (Besnard et al. 2011))

the tissue compartments, an index of lung sacculation and stromal thinning. Expression of a subset of mRNAs that encode proteins involved in surfactant homeostasis was highly correlated with increasing SatPC (surfactant lipid), body weight, lung weight, and structural maturation of the lung as gestation proceeded, representing a good example of integration gene expression profiling with lung physiology, biochemistry and morphological changes during lung maturation.

Application of Clustering Methods to In Vivo Models of Pulmonary Development "Phenoclustering," based on the integration of gene expression profiling data from genetic models used to study lung maturation, is useful in the identification of the biological processes mediating lung maturation. RNA profiling of lung from mouse models in which specific genes were deleted or expressed was useful in identification of transcription factors and signaling molecules in the respiratory epithelium that are critical for respiratory adaptation at birth, including NKX2-1, FOXA2, C/EBPα, MIA1 and CNB1. Conditional deletion or mutation of these transcription factors caused phenotypic and biochemical changes similar to those observed in respiratory distress syndrome (i.e., decreased surfactant production, lack of AT1 and AT2 cell differentiation, and inhibition of structural maturation) (Dave et al. 2006; DeFelice et al. 2003; Lin et al. 2008; Martis et al. 2006; Wan et al. 2004). Meta-analysis of RNA microarray datasets from this "phenocluster" showed that although these TFs and signaling molecules act through different signaling pathways and bind to distinct cis-elements, each regulates expression of shared transcriptional targets involved in surfactant protein and lipid biosynthesis (e.g., *Abca3, Scd1, Pon1, Sftpa, Sftpb, Sftpc and Sftpd*), fluid and solute transport (e.g., *Aqp5, Scnn1g, Slc34a2*) and innate host defense (e.g., *Lys, Sftpa, Sftpd and Scgb1a1*), indicating that Foxa2, CEBPα, Cnb1, Mia1 and TTF-1 may interact in a common transcriptional network regulating perinatal lung maturation and postnatal respiratory adaptation (Fig. 19.3a).

Based on the assumption that co-expressed genes sharing similar phenotypes are likely controlled by the same sets of transcription regulators, search of cis-elements shared by genes in the phenocluster was used to identify statistically overly-represented TFBSs. These TFBSs were ranked based on the total number of potential target genes containing each TFBS in promoter regions of genes in each cluster (Fig. 19.3b). Cytoscape v2.8.2 (http://www.cytoscape.org/) was used to generate a transcriptional regulatory network to visualize the predicted molecular interaction network linking this group of genes and predicted TFs. Nfatc3, Nkx2-1 and Foxa1/a2 were among the most overly-represented transcription factors in the lung with high connectivity to the genes comprising the network (Fig. 19.3b). The prediction was cross validated via the analysis of NKX2-1 ChIP-seq, there we scanned the presence of NFAT, CEBPA and FOXA2 binding sites within the peak regions containing NKX2-1 binding site using random sequence fragments as reference to calculate the binomial probability of the binding sites association. We identified that the co-binding probability and frequency of NKX2-1/CEBPA, NKX2-1/FOXA2 and NKX2-1/NFAT were significantly enriched in the positive NKX2-1 peak. These data support the concept that CEBPA, FOXA2 and NFAT act as interaction partners with NKX2-1 to regulate gene expression during lung development and maturation.



C

Matrix Family	TF Symbol	Binding Frequency
V$GATA	Gata1, Gata5, Gata6	42
V$HNF1	Hnf1a, Hnf1b	40
V$PAX5	Pax9	40
V$NFAT	Ilf3, Nfatc3	39
V$NKXH	Nkx2-1	35
V$AP1R	Bach1, Bach2	33
V$FKHD	Foxa1, Foxa2	32
V$SREB	Srebf1	28
V$ETSF	Gabpa, Gabpb1	24
V$IRFF	Irf1	8

Fig. 19.3 Meta-analysis of RNA microarray data from mouse models sharing common perinatal respiratory distress phenotypes. (**a**) Comparative RNA microarray data derived from lung epithelial selective deletion of *Foxa2*, *Cebpa*, *Cnb1*, mutation of *Nkx2-1*, and misexpression of *Mia1* identified common targets involved in surfactant biosynthesis, fluid/solute transport, and innate host defense. (**b**) Overly represented TFBS were identified using Clover and TRANSFAC software in the 2 kb promoter regions of the genes altered in all five microarrays. Cytoscape v2.8.2 was used to generate a Transcriptional Regulatory Network (TRN) via mapping of TF binding site to predicted target genes. If multiple TFs were associated with the same TF binding site, only TFs abundantly expressed in the lung were selected as representative TFs in the Table. *White rectangles* represent a TFBS family identified by a predicted consensus binding site and each *blue oval* represents a target gene

19.5.3 Data/Knowledge Integration

After the identification of major gene clusters, biologists often ask "What is unique about this gene set?" There are two common approaches to answering this question. A gene centric approach, identifies mRNAs of personal interest for further study, in contrast, an unbiased "systems biology" approach, can be used to identify themes, trends, and biological meaning implicit in the data. Biological knowledge and concepts integration represent an unbiased way to identify potentially important biological themes represented by distinct gene data sets and help in the assignment of potential roles of previously uncharacterized genes to biological processes.

Gene Set Enrichment Analysis Individual genes are associated with multiple biological annotations from various resources (Gene Ontology terms, Medical Subject Headings and keywords, pathways, protein–protein interactions, protein functional domains, phenotypes, literature/abstract etc.). Enrichment of genes in these functional categories can be determined using Fisher's exact test to compare the occurrence of terms in experimental gene sets, with annotations referenced in the rest of the genome. Thus overly represented functional categories can be identified in the gene list. Multiple pre-compiled web-based functional annotation tools including DAVID (Dennis et al. 2003), GSEA (Subramanian et al. 2005), and ToppGene (Chen et al. 2009b) are available that facilitate biologist to interpret high-throughput data in an unbiased way. Most of these methods compare functional representations within random gene lists or background genome to generate adjusted p-values that represent the statistical probability of observing an enriched functional category in experimental data sets. For genes within a cluster, Kappa similarity is a useful measure of functional similarity among genes based on the number of shared annotation terms (McGinn et al. 2004). Kappa similarity values range from 0 to 1, the higher the value of Kappa, the stronger the agreement in annotation terms.

RNA-Seq and ChIP-seq Integration: Chromatin immunoprecipitation (ChIP) followed by massive parallel sequencing (ChIP-seq) is a widely used approach for genome-wide identification of physical interactions between TFs and DNA sequence binding sites, thus providing evidence of potential physical binding (Ho et al. 2011). The integration of ChIP-seq data with mRNA expression profiling enables predictions of direct vs. indirect effects of transcription factors on target genes. Correlation of differentially expressed genes identified from RNA Sequence and/or microarray analyses with binding sites identified by ChIP-seq analysis support the likelihood but not prove the direct transcriptional relationship between the transcription factor and its target gene. Genes, whose expression altered in mRNA expression analysis, but not by ChIP-seq, are more likely to indicate secondary effects of perturbation of transcription factors. Enriched binding sites not associated with changes in mRNA expression are more likely to be nonfunctional binding sites or sites that are active only within distinct biological contexts. Because expression data and ChIP-seq data provide complementary information, their integration may enhance the accuracy of

predicting of transcription factor-target relationships than those based on single data sources.

Knowledge Integration Genome-wide biological knowledge bases have been established by public/private efforts focusing on the gene/protein centric functional annotation and curation; for example, NCBI Entrez Gene (http://www.ncbi.nlm.nih.gov/gene), Ensembl (http://useast.ensembl.org/index.html), GeneCards (http://www.genecards.org), ENCODE (https://www.encodeproject.org), UCSC genome browser (https://genome.ucsc.edu), UniProt Knowledgebase (http://www.uniprot.org/uniprot/), Protein Information Resource (PIR, http://pir.georgetown.edu), The Human Protein Atlas (http://www.proteinatlas.org), Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/), Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/omim), Human Gene Mutation Database (HGMD, http://www.hgmd.org) and Ingenuity knowledge base (IPA, http://www.ingenuity.com/products/ipa). These resources provide exceptional depth of knowledge linked to genes and proteins. Incorporation of knowledge derived from different genomic and biomedical information resources can enhance interpretation of "omics" data. Algorithms for integrating different types of data by combining high-throughput "omics" data with knowledge from both clinical and experimental observations are increasingly useful (see Table 19.1 for summary of genomic resources).

Network Modeling and Optimization A fundamental challenge in the "post genomic era" is to decipher the transcriptional regulatory mechanisms that direct intricate patterns of gene expression during organ formation. A wide range of methods has been developed to infer genome-scale regulatory networks from gene expression datasets (De Smet and Marchal 2010). Marbach et al performed a comprehensive assessment of over 30 network inference methods on microarray datasets from multiple resources (Marbach et al. 2012). They concluded that no single method performs optimally across all data sets. In contrast, integrative predictions from multiple inference methods showed robust and high performance across diverse data sets. Since methods of different categories (i.e., regression, mutual information, correlation, Bayesian networks) show intrinsic bias to the prediction of certain type of interactions, method integration provides complementary advantages and therefore is a powerful approach for optimization of regulatory networks. Network development and optimization can be achieved via multi-level integration (i.e., data, knowledge and method integration, Fig. 19.4a).

A consolidation pipeline was developed for the ensemble of network from different algorithms, integrating biological knowledge, expression data, and ChIP-seq data (Fig. 19.4a).

Driving Force Prediction The identification of essential regulators controlling cell fate and associated biological processes during organ development provides important insights into lung biology and serves to prioritize choices for experimental validation. We developed an algorithm to quantitatively measure node impor-

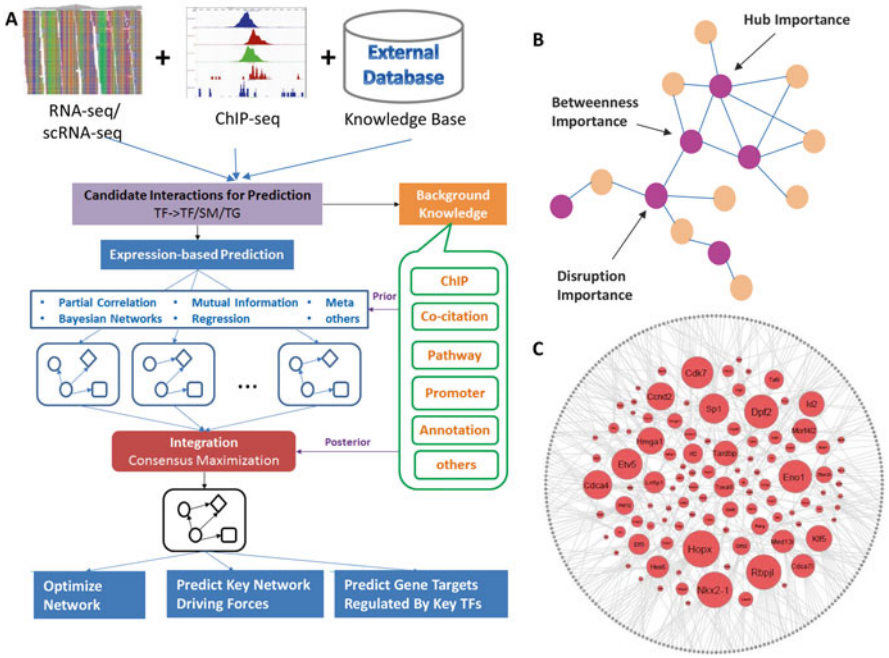


Fig. 19.4 Development of transcriptional regulatory network via data, method and knowledge integration. (**a**) The analytic strategy used to achieve consensus maximization of the predicted TRN. (**b**) Network schema illustrates ranking "driving forces" based their importance to the network. (**c**) A lung epithelial TRN was developed based on the RNA-seq of single cell isolated from E16.5 whole lung. Epithelial signature genes and TFs that are abundantly or selectively expressed in lung epithelial cells were used to generate the TF-TG network using a conditional dependency algorithm modified from G1DBN (Lebre 2009). Network contains total of 782 nodes and 1137 edges. Node size is proportional to the rank importance

tance in the network based on nodal centrality (determine the importance and nodal connectivity with all other nodes in a network) and/or disruption (determine how the removal of nodes in the network affects a network structure) (Borgatti et al. 2009; Hahn and Kern 2005). Three centrality metrics and three disruption metrics were used to calculate nodal importance are illustrated in Fig. 19.4b. (1). *Degree Centrality* measures the number of nodes that a given node is adjacent to is determined. A node with a high degree of centrality is one that can potentially influence many others (Hub); (2). *Closeness Centrality* measures the sum of geodesic distances from a given node to all others. A node with a low distance is likely to influence many others; (3). *Betweenness Centrality* measures the number of shortest paths that pass through a given node. A node with high betweenness is responsible for connecting many pairs of nodes via the best path; (4). *Disruptive Fragmentation* estimates the impact of the removal of a node on the disruption of the residual network; (5). *Disruptive Connection* assesses the impact of the removal of a node on

nodal connections in the residual network; and (6). *Disruptive Distance* estimates the impact of the removal of a node on the shortest path between nodes in the residual network. The rank sum score from these metrics is then used to rank order genes in the network. Figure 19.4c shows an example of prediction of essential driving forces in a lung epithelial cell gene network. A transcriptional regulatory network was developed based on the RNA-seq of single cell isolated from E16.5 whole lung using the conditional dependency algorithm modified from G1DBN (Lebre 2009). This network contains 782 nodes and 1137 edges. Node size is proportional to rank importance. *Nkx2-1, Hopx, Rbpjl, Etv5, Klf5* and *Id2* are among the top ranked important driving forces for lung epithelial development and differentiation at E16.5 (Fig. 19.4c).

19.5.4 Meta-Analyses of "omics" Data

As a result of 'omics' research, high-throughput data sets available to the research community for further analysis are rapidly growing. "Meta-analysis" uses statistical tools that combine results from several related, but independent experiments (Hong and Breitling 2008). Increasing statistical power enabling detection of treatment effects, assessment of the variability among studies, and maximizes the use of available data. To identify mechanisms by which SREBP regulates lung lipid homeostasis, we designed a meta-analysis of available genome-wide RNA expression and ChIP-seq datasets related to SCAP/SREBP/INSIG in lung and liver. Eighty experimental data sets were collected from GEO (http://www.ncbi.nlm.nih.gov/geo/). Through the comprehensive data analysis, we identified general (i.e., common in liver and lung) vs. lung specific effect of SREBP signaling and associated genes, which in turn, lead the development of lung specific vs. general SREBP-related gene regulatory networks. Through the network structural centrality and disruption measurements, we identified DBP, NR5A1, PPARG, RARA, and STAT5A as important driving forces in the general and EGR1, CEBPA/B/D, ATF3, FOXA2, and SOX9 in lung specific SREBP-centered gene networks (Fig. 19.5a, b).

19.6 Developing a Systems Level Understanding of Surfactant Homeostasis and Lung Maturation

We have introduced a number of commonly used bioinformatics approaches and their applications in the study of lung maturation. One thing worth noting is that no single method is sufficient and without limitation. For example, GO annotations and literature mining enable the association of genes with biological processes and pathways, but are limited to and biased by current knowledge. While transcription factor and target gene correlations take into account that expression profiles of

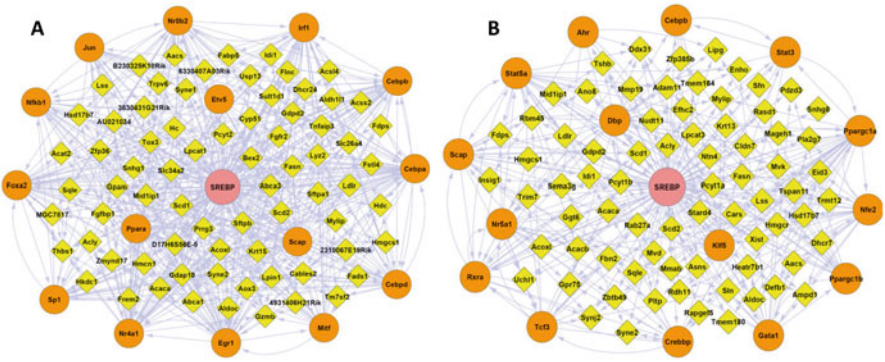


Fig. 19.5 SREBP centered TRN from meta-analyses of RNA microarray, RNA-seq, and ChIP-seq data. Networks were developed via meta-analysis of available genome-wide RNA expression and ChIP-seq datasets related to SCAP/SREBP/INSIG in lung and liver. Eighty experimental data sets were collected from GEO (http://www.ncbi.nlm.nih.gov/geo/). General (i.e., common in liver and lung) vs. lung specific effect of SREBP signaling and associated genes were identified, which in turn, lead the development of (**a**) lung specific vs. (**b**) general SREBP-related gene regulatory networks. Regulatory interactions between TF and target genes were predicted using partial correlation inferences (Opgen-Rhein and Strimmer 2007)

transcription factors and their targets are often correlated, many transcription factors regulate gene expression via post-transcriptional mechanisms, including transcript stability, DNA binding site accessibility, interactions with tissue-specific co-factors and dynamic changes in chromatin structure. Promoter analysis seeking to identify conserved TFBSs in promoters of co-expressed genes is useful in identifying potential cis-elements, but does not identify the binding or activity of transcription factors. The frequency and degenerate nature of many TFBS motifs contributes to the complexity of promoter/enhancer predictions. Transgenic animal models with the deletion or over-expression of essential transcriptional regulators are useful in identifying downstream targets and inferring regulatory interactions. However, experiments may be confounded by non-biological/drastic perturbations of genetic networks outside the range of physiological relevance. Direct and indirect transcriptional targets are not distinguished by expression profiling approach. By applying systems biology approaches; one can take the full advantage of the fast growing data and available methodology resources, to integrate related data from diverse technical platforms to achieve a more comprehensive view of organ development and function.

19.6.1 Regulatory Networks Regulating Lung Surfactant Homeostasis: Static Model

Pulmonary surfactant is required for lung function at birth and throughout life. Lung lipid and surfactant homeostasis requires regulation of multi-tiered processes, coordinating the synthesis, storage, secretion, reuptake, and degradation of surfactant proteins and lipids by AT2 cells. To generate a transcriptional model by which

surfactant homeostasis is controlled, we retrieved 194 mRNA microarray samples from 27 distinct mouse models in which transcription factors /signaling molecules modifications were made in mouse models of lung disease. An algorithm integrating expression profiling with expression-independent knowledge using Gene Ontology (GO) similarity analysis, promoter (gene regulatory sequences) motif, searching, protein-protein interactions and literature mining was developed to model genetic networks regulating surfactant homeostasis and related biological processes in lung. A transcription factor - target gene similarity matrix was generated by integrating data derived from different analytic platforms. A confidence scoring function was built to rank the likelihood of a transcription factor regulating its target gene. Using this strategy, critical components of transcriptional networks directing lipogenesis, lipid trafficking, and surfactant homeostasis were predicted. SREBP, HNF3, ETSF, CEBP, GATA and IRFF were predicted as key regulatory hubs in this network, SREBP, FOXA2 and CEBPA together form a common core regulatory module that is predicted to control surfactant lipid homeostasis (Xu et al. 2010).

19.6.2 Transcriptional Programs Controlling Perinatal Lung Maturation (Dynamic Model)

The timing of lung maturation is likely controlled by complex genetic programs that are influenced by multiple environmental and temporal factors. Cross-sectional integrative gene expression profiling analysis does not take into account the dynamic nature of the transcriptional programs accompanying lung maturation. We used genetic and bioinformatics approaches to elucidate the relationship between the length of gestation and lung function at birth in two inbred mouse strains whose gestational length differed by 30 h (C57BL/6 J: 19.5 days and A/J: 21 days) (Besnard et al. 2011; Murray et al. 2010; Xu et al. 2012a). Shorter gestation in C57BL/6 J mice was associated with advanced morphological and biochemical pulmonary development and better perinatal survival compared to A/J pups born prematurely at 19.5 days of gestational age (Besnard et al. 2011). Developmental changes in lung RNAs in the two mouse strains were assessed using Mouse Gene 1.0 ST Arrays. A functional Bayesian approach (Angelini et al. 2008) was used to analyze time dependent changes in lung mRNA expression from each mouse strain. Matched temporal dependent expression patterns of transcription factors and targets during lung maturation were identified using STEM (Short Time-series Expression Miner), a clustering algorithm designed for analysis of short time series gene expression datasets (5–8 time points) (Ernst and Bar-Joseph 2006). Comprehensive knowledge integration was employed to identify biological processes underling the dynamic gene expression patterns during lung maturation. Both temporal and strain dependent genes and pathways were identified (Fig. 19.6).

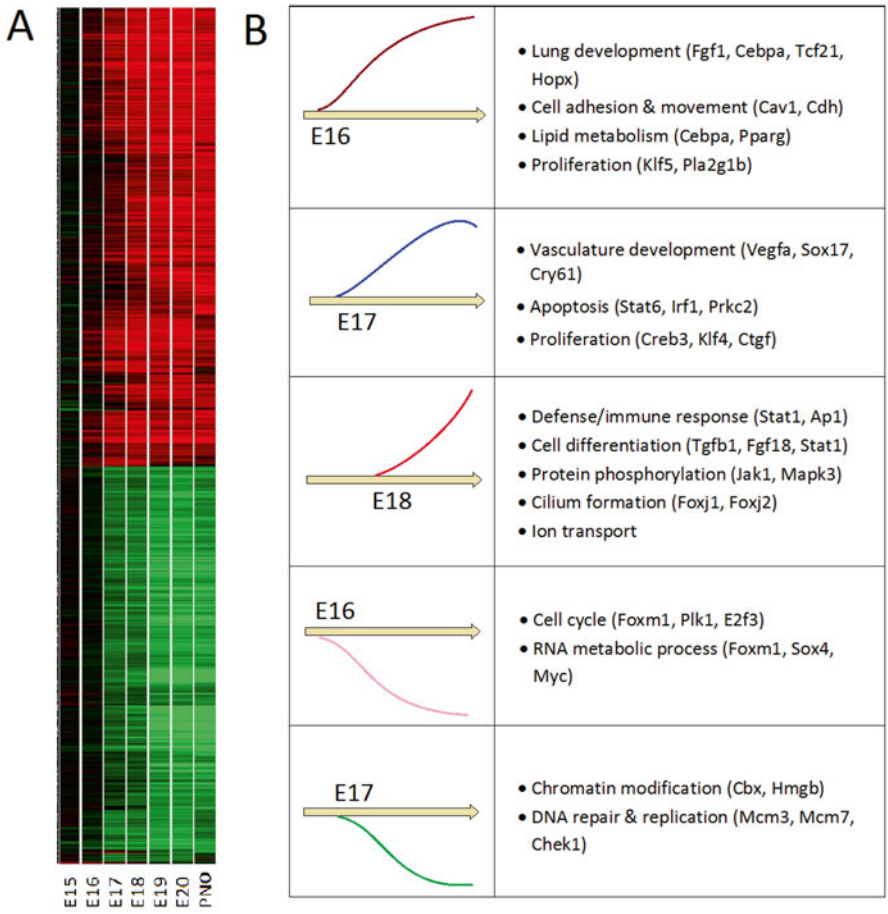


Fig. 19.6 Dynamic changes in gene expression during perinatal lung maturation. (**a**) Heatmap of temporal dependent gene expression changes during lung maturation (E15-PN0). (**b**) Schematized depiction of bioprocesses and predicted key regulators that change dynamically with advancing gestation is shown

Major bioprocesses and key regulators associated with different stages of lung development included "Lung development," "cell adhesion" and "cell movement," "lipid metabolism," and "proliferation" all induced early in lung maturation (E15-16 in the pseudoglandular stage). *Hopx, Cebpa, Tcf21* and *Klf5* were predicted to be important transcriptional regulators at this stage, a finding consistent with gene deletion studies that demonstrated their important roles in prenatal lung maturation and function (Martis et al. 2006; Quaggin et al. 1999; Wan et al. 2008; Yin et al.

2006). transcription factors regulating "vasculature development" and "apoptosis" were induced at E16-17 (canalicular stage). *Vegfa, Sox17 and Stat3/6* represented important regulators at this time. "Innate defense/immune responses," "cell differentiation," "protein phosphorylation," "ion transport," and "cilium formation" were induced at later gestational ages (E18-20, in the saccular stage). *Stat1, Tgfb1, and Foxj1* were identified as important regulators associated with the saccular stage of maturation. Cell cycle and chromatin assembly were progressively repressed during perinatal lung maturation. FOXM1, PLK1, chromobox, SWI/SNF and high mobility group families of TFs were predicted to play important roles in the negative regulation of cell proliferation that occurs before birth. Innate immune responses and surfactant production were connected processes necessary for respiration and survival after birth. In contrast, "epigenetic regulators" ware predicted to play a repressive role by altering chromatin structure and controlling the cell cycle (Xu et al. 2012b). We hypothesize that the precise regulation and balance among these gene networks serve to coordinate the timing of lung maturation with gestational length that differs in the B6 and A/J mouse strains as well as across terrestrial vertebrates.

19.6.3 Sub-Networks Control Distinct Biological Processes During Lung Maturation

Transcriptional Regulatory Networks (TRN) can be divided into sub-networks of interconnected genes, each representing distinct functional units within the entire network. Each distinct unit may be driven by tissue or cell type specific transcription factors /signaling molecules hubs and activates at specific times and in specific cell types. All functional units work coordinately to control temporal-spatial processes mediating lung maturation. Due to the complexity of interacting TRNs, it is useful to identify sub-networks that consist of smaller groupings of effector genes, usually centered within one or several interrelated hubs. Effector genes in the sub-network tend to be co-expressed, transcriptionally co-regulated and perform similar cellular functions or work in concert to influence specific developmental processes. Future experimental validation of predicted hubs and effector genes in sub-networks will be useful in identifying the biological processes involved in lung maturation. An example of an NKX2-1 related sub-network, consisting of potential TF partners or cofactors of NKX2-1 (pink nodes) and predicted gene targets of NKX2-1 (blue nodes) is shown in Fig. 19.7c.

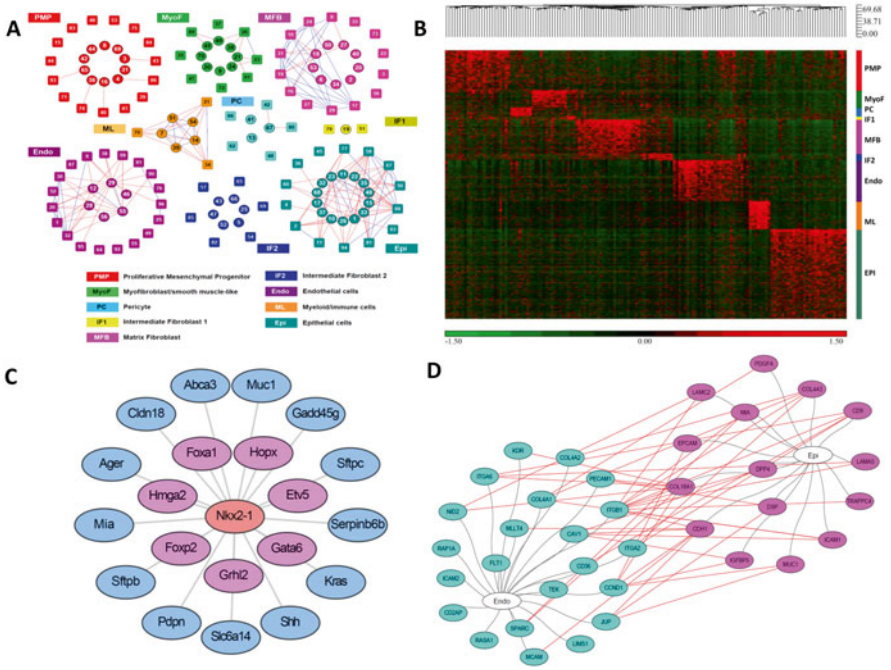


Fig. 19.7 scRNA-seq from E16.5 whole lung was analyzed using the SINCERA analytic pipeline. (**a**) Hierarchical clustering of 148 cells predicted nine distinct pulmonary cell types. Average linkage was used to calculate cluster related cells using centered Pearson correlations for similarity measurement. Minimum similarity was set as 0.5. Each color represents a distinct cell cluster. Cells in the outer and inner wheels represent single cells isolated from two distinct experiments. Connecting lines indicated the z-score correlation between each node >= 0.05. (**b**) Heatmap overview of the 2D clustering of signature RNAs from 148 single cells (E16.5). Signature genes were selected based on FPKM, expression specificity index, and t-Test p-value. (**c**) The predicted NKX2-1 sub-networks (1-hop) was extracted from the epithelial cell TRN at E16.5. Epithelial "signature" genes and TFs that were abundantly or selectively expressed in lung epithelial cells were used to generate a TF-TG network using a conditional dependency algorithm modified from G1DBN (Lebre 2009). Pink nodes are potential TF partners or cofactors of NKX2-1, blue nodes are predicted gene targets of NKX2-1. (**d**) Cross talk between lung epithelial cells and endothelial cells, via extracellular matrix protein-protein interactions is shown. Genes encoding ECM proteins (according to gene ontology annotation) were extracted from the endothelial (*yellow nodes*) and epithelial (*blue nodes*) cell clusters. *Edges* represent predicted protein-protein interactions

19.7 Understanding Lung Maturation at Single Cell Level

Historically, gene expression profiling was performed on populations of cells or in whole organs. Bulk measurements of cell populations or mixed populations may be obscured by changes in cellular composition during organ growth and differentiation. The role of rare subsets of cells may be lost, and the interplay among different cell types covered in analyses of whole tissue containing multiple cell types. Advances in single-cell genomics and proteomics enable discovery of previously

undetected subpopulations of cells, lineage relationships, and mechanisms by which individual cells interact during organogenesis.

While the future for single-cell next-generation sequencing based genomic studies is promising, it brings new and specific analytical challenges. Most of the previous available methods were designed for quantifying the mean behaviors of millions of cells by averaging the signal of individual cells. Although some tools for analyzing RNA-seq and microarray data from bulk cell populations can be applied to scRNA-seq data, new analytic strategies and workflows are required to address the unique issues and opportunities associated with single cell data, including the identification and characterization of unknown cell types, handing confounding factors such as batch and cell cycle effects, addressing the cellular heterogeneity in complex biological systems, to name a few (Brennecke et al. 2013; Buettner et al. 2015; Kharchenko et al. 2014; Kim and Marioni 2013; Trapnell et al. 2014). For cell type identification, most single cell studies used hierarchical clustering or PCA-like methods or the combination of the two (Darmanis et al. 2015; Treutlein et al. 2014; Usoskin et al. 2015; Xue et al. 2013; Yan et al. 2013). Recently, a number of methods specifically designed for scRNA-seq analysis have been introduced including SNN-Cliq (Xu and Su 2015), scLVM (Buettner et al. 2015) and BackSPIN (Zeisel et al. 2015) for clustering; SAMstrt and Bayesian approach for single-cell differential expression analysis (Katayama et al. 2013; Kharchenko et al. 2014; Li and Tibshirani 2013); SINGuLAR Analysis Toolset (https://cn.fluidigm.com/software) enabling identification of genes that are either differentially expressed or co-expressed; ICGS (*Iterative Clustering and Guide-Gene Selection* https://code.google.com/p/altanalyze/) for unbiased identification of the most coherent, correlated gene signatures that are able to segregate distinct developmental states or cell-types; Monocle (Trapnell et al. 2014) and SCUBA (Marco et al. 2014) for extracting lineage relationships from scRNA-seq and modeling the dynamic changes associated with cell differentiation. Recent studies by (Satija et al. 2015; Pettit et al. 2014) combined single-cell RNA-seq gene expression profiles with complementary in situ hybridization (ISH) data to reveal the 3D expression patterns. These efforts addressed spatial localization more directly and precisely than previous efforts using independent component analysis (ICA) or principal component analysis (PCA) to approximate spatial location. These advanced methods mostly focused on one aspect of the data analysis. How to design the analytic workflow to process large amounts scRNA-seq data from heterogeneous cell populations and reveal biological insights represent a substantial challenge for most investigators.

We developed SINCERA, a computational pipeline for SINgle CEll RNA-seq profiling Analysis, that enables investigators analyzing scRNA-seq data using standard desktop/laptop computers to conduct data filtering, normalization, clustering, cell type identification, gene signature prediction, TRN construction, and identification of driving force (key nodes) for each cell type (Guo et al. 2015). We applied SINCERA pipeline to single cells from the embryonic mouse lung at the saccular phase (E16.5/E18.5) of morphogenesis, to identify major cell types including epithelial, fibroblast, endothelial, myo-fibroblast/smooth muscle, pericyte and myeloid cells. Gene signatures, surface markers, bioprocesses and

functional profiles associated with each cell type were predicted, Fig. 19.7a, b. SINCERA provides an analytic tool to delineate cell type specific regulatory networks, key regulators, and predict cell-cell communications across cell types via paracrine and autocrine signaling and protein-protein interactions, Fig. 19.7c, d. We used network-based approaches to identify cell type specific transcriptional networks and key regulators within the networks. As shown in Fig. 19.7c, epithelial signature genes and transcriptional factors that are abundantly or selectively expressed in lung epithelial cells were used to generate a transcriptional network using conditional dependency algorithm modified from G1DBN (Lebre 2009). NKX2-1 was predicted as one of the top ranked transcriptional factors at E16.5, Fig. 19.4c. The NKX2-1 sub-network was extracted from a global epithelial cell transcriptional network, Fig. 19.7c. Using cell type specific gene signatures and pre-compiled protein interaction databases, the shortest path between cell types (e.g., shortest path between endothelial and epithelial cells) was calculated from a protein-protein interaction and association network, Fig. 19.7d. For example, the network predicts selective interactions between endothelial, Jup expressing, and epithelial, Cdh1 expressing cells. *Jup* (junction plakoglobin), expressed by endothelial cells, functions as a substrate for vascular endothelial protein tyrosine phosphatase. Jup forms adhesion complexes with E-cadherin (Cdh1) that regulates *Jup* expression (Auersperg et al. 1999; Luo et al. 1999), Fig. 19.7d. The finding that experimentally validated protein-protein interactions between endothelial and epithelial cells were identified at the single cell level support the potential utility of predicting "social networks" among distinct cell types using single cell transcriptomics.

To facilitate access to single cell data and its applications for study of lung development, we developed "LungGENS" (Lung Gene Expression iN Single-cell), a web tool useful for mapping single cell gene expression (Du et al. 2015). The current version of LungGENS supports three major functions: 'Query by single gene,' 'Query by gene list,' and 'Query by cell type.' 'Query by single gene' provides quantitative RNA expression of the gene of interest in each lung cell type. 'Query by gene list' enables the user to input their list of genes of interest to identify the cell type selectively expressing in that gene set. 'Query by cell type' returns selective gene signatures and genes encoding cell surface markers and transcriptional factors via interactive heatmaps and tables. LungGENS serves as a rich knowledge base and is broadly applicable for lung research, providing a cell-specific RNA expression resource at single-cell resolution. We are actively developing a new version of LungGENS with extended the scope and contents of the database to include lung developmental studies from single cell, sorted cell populations and whole lung tissues. Both SINCERA (https://research.cchmc.org/pbge/sincera.html) and LungGENS (https://research.cchmc.org/pbge/lunggens/default.html) are freely available to the public providing valuable tools for single cell transcriptomic analysis. Understanding the cellular and molecular mechanisms driving normal lung function and formation will enhance our understanding of the pathogeneses of lung diseases affecting infants and children.

19.8 Conclusions

Integration of increasingly complex structural, temporal, and genetic data with the cellular processes that determine lung maturation and function in the perinatal period remain a formidable challenge. Lung maturation requires precise temporal and spatial regulation of multiple signaling and transcriptional events among a great diversity of lung cell types. Identification of lung cell progenitors, cell lineage, and fate determination orchestrating the formation and function of the saccular-alveolar unit is dependent upon a myriad of transcriptional and signaling networks active in each cell and the interactions among both neighboring and distant cells. These challenges are not fully met without systematic analysis of the transcriptional regulatory networks controlling lung organogenesis and function. In this chapter, we have summarized bioinformatic and systems biology approaches being applied to the study of perinatal lung maturation and surfactant homeostasis. We emphasize the application of integrative approaches to achieve a more comprehensive understanding of lung maturation at a systems biology level. The transcriptional factors participating in lung development are not exclusively lung-specific and are used by other organs to regulate organogenesis and organ function. The unique combinations and interactions among transcriptional factors/signaling molecules and interactions among distinct cell types are likely to mediate the structural and functional specificities that mediate lung formation and maturation. Identification of key transcriptional factors, cofactors, and signature genes in individual pulmonary cell types at different stages of lung development will be needed to further understand the precise temporal sequences and dynamic spatial changes occurring during formation of the lung. Present single cell findings predict complex cell communications via direct cell-cell contact, cell-matrix interactions, and through paracrine and autocrine cell signalling that coordinate the formation and maturation of the peripheral lung. A thorough understanding of the molecular mechanisms controlling the coordination of length of gestation, with the timing of lung maturation will provide new insights useful for developing therapeutic and diagnostic tools for treatment and management of pulmonary diseases affecting children and adults. The systems biology strategies outlined in this review are highly relevant to studies related to other organs and diseases. Since pulmonary disease affects organ structure and function, single gene approaches are less likely to identify the mechanisms underlying the pathogenesis of lung disease. Systems biology, integrating multiscaled data at molecular, cellular, and organ levels will be useful in understanding disease pathogenesis. Open access to data provided by RNA and DNA sequencing, genomics, proteomics, metabolomics, and lipidomics, and the ability to integrate and interpret the increasingly complex data will play an increasingly important role in biology research and medicine.

References

Amos WB, White JG. How the confocal laser scanning microscope entered biological research. Biol Cell. 2003;95:335–42.

Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.

Angelini C, Cutillo L, De Canditiis D, Mutarelli M, Pensky M. BATS: a Bayesian user-friendly software for analyzing time series microarray experiments. BMC Bioinformatics. 2008;9:415.

Auersperg N, Pan J, Grove BD, Peterson T, Fisher J, Maines-Bandiera S, Somasiri A, Roskelley CD. E-cadherin induces mesenchymal-to-epithelial transition in human ovarian surface epithelium. Proc Natl Acad Sci U S A. 1999;96:6249–54.

Avery ME, Mead J. Surface properties in relation to atelectasis and hyaline membrane disease. AMA. 1959;97:517–23.

Bell SM, Zhang L, Mendell A, Xu Y, Haitchi HM, Lessard JL, Whitsett JA. Kruppel-like factor 5 is required for formation and differentiation of the bladder urothelium. Dev Biol. 2011;358:79–90.

Bell SM, Zhang L, Xu Y, Besnard V, Wert SE, Shroyer N, Whitsett JA. Kruppel-like factor 5 controls villus formation and initiation of cytodifferentiation in the embryonic intestinal epithelium. Dev Biol. 2013;375:128–39.

Besnard V, Xu Y, Whitsett JA. Sterol response element binding protein and thyroid transcription factor-1 (Nkx2.1) regulate Abca3 gene expression. Am J Physiol. 2007;293:L1395–405.

Besnard V, Wert SE, Stahlman MT, Postle AD, Xu Y, Ikegami M, Whitsett JA. Deletion of Scap in alveolar type II cells influences lung lipid homeostasis and identifies a compensatory role for pulmonary lipofibroblasts. J Biol Chem. 2009;284:4018–30.

Besnard V, Wert SE, Ikegami M, Xu Y, Heffner C, Murray SA, Donahue LR, Whitsett JA. Maternal synchronization of gestational length and lung maturation. PLoS One. 2011;6:e26682.

Bintu L, Yong J, Antebi YE, McCue K, Kazuki Y, Uno N, Oshimura M, Elowitz MB. Dynamics of epigenetic regulation at the single-cell level. Science. 2016;351:720–4.

Borgatti SP, Mehra A, Brass DJ, Labianca G. Network analysis in the social sciences. Science. 2009;323:892–5.

Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10:1093–5.

Bridges JP, Schehr A, Wang Y, Huo L, Besnard V, Ikegami M, Whitsett JA, Xu Y. Epithelial SCAP/INSIG/SREBP signaling regulates multiple biological processes during perinatal lung maturation. PLoS One. 2014;9, e91376.

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015;33:155–60.

Bunyavanich S, Schadt EE. Systems biology of asthma and allergic diseases: a multiscale approach. J Allergy Clin Immunol. 2015;135:31–42.

Burgess DJ. RNA: Putting transcriptomics in its place. Nat Rev Genet. 2015;16:319.

Burri PH. Fetal and postnatal development of the lung. Annu Rev Physiol. 1984;46:617–28.

Chen G, Korfhagen TR, Xu Y, Kitzmiller J, Wert SE, Maeda Y, Gregorieff A, Clevers H, Whitsett JA. SPDEF is required for mouse pulmonary goblet cell differentiation and regulates a network of genes associated with mucus production. J Clin Invest. 2009a;119(10):2914–24.

Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009b;37:W305–11.

Chen G, Wan H, Luo F, Zhang L, Xu Y, Lewkowich I, Wills-Karp M, Whitsett JA. Foxa2 programs Th2 cell-mediated innate immunity in the developing lung. J Immunol. 2010;184:6133–41.

Chen G, Korfhagen TR, Karp CL, Impey S, Xu Y, Randell SH, Kitzmiller J, Maeda Y, Haitchi HM, Sridharan A, et al. Foxa3 induces goblet cell metaplasia and inhibits innate antiviral immunity. Am J Respir Crit Care Med. 2014;189:301–13.

Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, Hayden Gephart MG, Barres BA, Quake SR. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci U S A. 2015;112:7285–90.

Dave V, Childs T, Xu Y, Ikegami M, Besnard V, Maeda Y, Wert SE, Neilson JR, Crabtree GR, Whitsett JA. Calcineurin/Nfat signaling is required for perinatal lung maturation and function. J Clin Invest. 2006;116:2597–609.

De Smet R, Marchal K. Advantages and limitations of current network inference methods. Nat Rev Microbiol. 2010;8:717–29.

DeFelice M, Silberschmidt D, DiLauro R, Xu Y, Wert SE, Weaver TE, Bachurski CJ, Clark JC, Whitsett JA. TTF-1 phosphorylation is required for peripheral lung morphogenesis, perinatal survival, and tissue-specific gene expression. J Biol Chem. 2003;278:35574–83.

Deng Q, Ramskold D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 2014;343:193–6.

Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol. 2003;4:P3.

Du Y, Guo M, Whitsett JA, Xu Y. 'LungGENS': a web-based tool for mapping single-cell gene expression in the developing lung. Thorax. 2015;70:1092–4.

Dubin SB. Assessment of fetal lung maturity: in search of the Holy Grail. Clin Chem. 1990;36:1867–9.

Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. BMC Bioinform. 2006;7:191.

Fu L, Medico E. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinform. 2007;8:3.

Gasch AP, Eisen MB. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Genome Biol. 2002;3. RESEARCH0059.1– RESEARCH0059.22.

Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. Lancet. 2008;371:75–84.

Gravett MG, Rubens CE, Nunes TM. Global report on preterm birth and stillbirth (2 of 7): discovery science. BMC pregnancy and childbirth. 2010;10(1):S2.

Grenache DG, Gronowski AM. Fetal lung maturity. Clin Biochem. 2006;39:1–10.

Guo G, Huss M, Tong GQ, Wang C, Li Sun L, Clarke ND, Robson P. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev Cell. 2010;18:675–85.

Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. PLoS Comput Biol. 2015;11:e1004575.

Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol. 2005;22:803–6.

Hitzemann R, Bottomly D, Darakjian P, Walter N, Iancu O, Searles R, Wilmot B, McWeeney S. Genes, behavior and next-generation RNA sequencing. Genes Brain Behav. 2013;12:1–12.

Ho JW, Bishop E, Karchenko PV, Negre N, White KP, Park PJ. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. BMC Genomics. 2011;12:134.

Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci U S A. 2000;97:8409–14.

Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics. 2008;24:374–82.

Inerot S, Heinegard D, Olsson SE, Telhag H, Audell L. Proteoglycan alterations during developing experimental osteoarthritis in a novel hip joint model. J Orthop Res. 1991;9:658–73.

Jaenisch R, Young R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell. 2008;132:567–82.

Katayama S, Tohonen V, Linnarsson S, Kere J. SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. Bioinformatics. 2013;29:2943–5.

Kelly AD, Hill KE, Correll M, Hu L, Wang YE, Rubio R, Duan S, Quackenbush J, Spentzos D. Next-generation sequencing and microarray-based interrogation of microRNAs from formalin-fixed, paraffin-embedded tissue: preliminary assessment of cross-platform concordance. Genomics. 2013;102:8–14.

Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11:740–2.

Kherlopian AR, Song T, Duan Q, Neimark MA, Po MJ, Gohagan JK, Laine AF. A review of imaging techniques for systems biology. BMC Syst Biol. 2008;2:74.

Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol. 2013;14:R7.

Kitano H. Computational systems biology. Nature. 2002;420:206–10.

Lange AW, Haitchi HM, Lecras TD, Sridharan A, Xu Y, Wert SE, James J, Udell N, Thurner PJ, Whitsett JA. Sox17 is required for normal pulmonary vascular morphogenesis. Dev Biol. 2014;387:109–20.

Lebre S. Inferring dynamic genetic networks with low order independencies. Stat Appl Genet Mol Biol. 2009;8. Article 9.

Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013;22:519–36.

Lin S, Ikegami M, Xu Y, Bosserhoff AK, Malkinson AM, Shannon JM. Misexpression of MIA disrupts lung morphogenesis and causes neonatal death. Dev Biol. 2008;316:441–55.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

Luo J, Lubaroff DM, Hendrix MJ. Suppression of prostate cancer invasive potential and matrix metalloproteinase activity by E-cadherin transfection. Cancer Res. 1999;59:3552–6.

Maeda Y, Dave V, Whitsett JA. Transcriptional control of lung morphogenesis. Physiol Rev. 2007;87:219–44.

Maeda Y, Chen G, Xu Y, Haitchi HM, Du L, Keiser AR, Howarth PH, Davies DE, Holgate ST, Whitsett JA. Airway epithelial transcription factor NK2 homeobox 1 inhibits mucous cell metaplasia and Th2 inflammation. Am J Respir Crit Care Med. 2011;184:421–9.

Maeda Y, Tsuchiya T, Hao H, Tompkins DH, Xu Y, Mucenski ML, Du L, Keiser AR, Fukazawa T, Naomoto Y, et al. Kras(G12D) and Nkx2-1 haploinsufficiency induce mucinous adenocarcinoma of the lung. J Clin Invest. 2012;122:4388–400.

Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9:796–804.

Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan GC. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc Natl Acad Sci U S A. 2014;111:E5643–50.

Martis PC, Whitsett JA, Xu Y, Perl AK, Wan H, Ikegami M. C/EBPalpha is required for lung maturation at birth. Development. 2006;133:1155–64.

Mayburd AL, Martlinez A, Sackett D, Liu H, Shih J, Tauler J, Avis I, Mulshine JL. Ingenuity network-assisted transcription profiling: Identification of a new pharmacologic mechanism for MK886. Clin Cancer Res. 2006;12:1820–7.

McGinn T, Wyer PC, Newman TB, Keitz S, Leipzig R, For GG. Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). CMAJ. 2004;171:1369–73.

McMurtry IF. Introduction: pre- and postnatal lung development, maturation, and plasticity. Am J Physiol. 2002;282:L341–4.

Metzger DE, Xu Y, Shannon JM. Elf5 is an epithelium-specific, fibroblast growth factor-sensitive transcription factor in the embryonic lung. Dev Dyn. 2007;236:1175–92.

Metzger RJ, Klein OD, Martin GR, Krasnow MA. The branching programme of mouse lung development. Nature. 2008;453:745–50.

Miller LA, Wert SE, Clark JC, Xu Y, Perl AK, Whitsett JA. Role of Sonic hedgehog in patterning of tracheal-bronchial cartilage and the peripheral lung. Dev Dyn. 2004;231:57–71.

Minn AJ, Gupta GP, Siegel PM, Bos PD, Shu W, Giri DD, Viale A, Olshen AB, Gerald WL, Massague J. Genes that mediate breast cancer metastasis to lung. Nature. 2005;436:518–24.
Muglia LJ, Katz M. The enigma of spontaneous preterm birth. N Engl J Med. 2010;362:529–35.
Murray SA, Morgan JL, Kane C, Sharma Y, Heffner CS, Lake J, Donahue LR. Mouse gestation length is genetically determined. PLoS One. 2010;5:e12418.
Nikolayeva O, Robinson MD. edgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology. Methods Mol Biol. 2014;1150:45–79.
Opgen-Rhein R, Strimmer K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst Biol. 2007;1:37.
Perl AK, Whitsett JA. Molecular mechanisms controlling lung morphogenesis. Clin Genet. 1999;56:14–27.
Pettit JB, Tomer R, Achim K, Richardson S, Azizi L, Marioni J. Identifying cell types from spatially referenced single-cell expression datasets. PLoS Comput Biol. 2014;10:e1003824.
Qiu S, Luo S, Evgrafov O, Li R, Schroth GP, Levitt P, Knowles JA, Wang K. Single-neuron RNA-Seq: technical feasibility and reproducibility. Front Genet. 2012;3:124.
Quaggin SE, Schwartz L, Cui S, Igarashi P, Deimling J, Post M, Rossant J. The basic-helix-loop-helix protein pod1 is critically important for kidney and lung organogenesis. Development. 1999;126:5771–83.
Rajavelu P, Chen G, Xu Y, Kitzmiller JA, Korfhagen TR, Whitsett JA. Airway epithelial SPDEF integrates goblet cell differentiation and pulmonary Th2 inflammation. J Clin Invest. 2015;125:2021–31.
Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30:777–82.
Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. 2014;42:8845–60.
Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33:495–502.
Seo J, Gordish-Dressman H, Hoffman EP. An interactive power analysis tool for microarray hypothesis testing and generation. Bioinformatics. 2006;22:808–14.
Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet. 2013;14:618–30.
St Croix CM, Shand SH, Watkins SC. Confocal microscopy: comparisons, applications, and problems. Bio Tech. 2005;39:S2–5.
Stelzer Y, Shivalila CS, Soldner F, Markoulaki S, Jaenisch R. Tracing dynamic changes of DNA methylation at single-cell resolution. Cell. 2015;163:218–29.
Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6:377–82.
Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K, Surani MA. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell. 2010;6:468–78.
Ten Have-Opbroek AA. Lung development in the mouse embryo. Exp Lung Res. 1991;17:111–30.
Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32:381–6.
Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 2014;509:371–5.

Tseng GC, Wong WH. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. Biometrics. 2005;61:10–6.
Usoskin D, Furlan A, Islam S, Abdo H, Lonnerberg P, Lou D, Hjerling-Leffler J, Haeggstrom J, Kharchenko O, Kharchenko PV, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci. 2015;18:145–53.
Vaughan AE, Brumwell AN, Xi Y, Gotts JE, Brownfield DG, Treutlein B, Tan K, Tan V, Liu FC, Looney MR, et al. Lineage-negative progenitors mobilize to regenerate lung epithelium after major injury. Nature. 2015;517:621–5.
Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett Jr DE, Hieter P, Vogelstein B, Kinzler KW. Characterization of the yeast transcriptome. Cell. 1997;88:243–51.
Vielreicher M, Schurmann S, Detsch R, Schmidt MA, Buttgereit A, Boccaccini A, Friedrich O. Taking a deep look: modern microscopy technologies to optimize the design and functionality of biocompatible scaffolds for tissue engineering in regenerative medicine. J R Soc Interface / R Soc. 2013;10:20130263.
Vince JE, Wong WW, Khan N, Feltham R, Chau D, Ahmed AU, Benetatos CA, Chunduru SK, Condon SM, McKinlay M, et al. IAP antagonists target cIAP1 to induce TNFalpha-dependent apoptosis. Cell. 2007;131:682–93.
Wan H, Xu Y, Ikegami M, Stahlman MT, Kaestner KH, Ang SL, Whitsett JA. Foxa2 is required for transition to air breathing at birth. Proc Natl Acad Sci U S A. 2004;101:14449–54.
Wan H, Dingle S, Xu Y, Besnard V, Kaestner KH, Ang SL, Wert S, Stahlman MT, Whitsett JA. Compensatory roles of Foxa1 and Foxa2 during lung morphogenesis. J Biol Chem. 2005;280:13809–16.
Wan H, Luo F, Wert SE, Zhang L, Xu Y, Ikegami M, Maeda Y, Bell SM, Whitsett JA. Kruppel-like factor 5 is required for perinatal lung morphogenesis and function. Development. 2008;135:2563–72.
Wang T, Hopkins D, Schmidt C, Silva S, Houghton R, Takita H, Repasky E, Reed SG. Identification of genes differentially over-expressed in lung squamous cell carcinoma using combination of cDNA subtraction and microarray analysis. Oncogene. 2000;19:1519–28.
Weaver TE, Beck DC. Use of knockout mice to study surfactant protein structure and function. Biol Neonate. 1999;76 Suppl 1:15–8.
Weaver TE, Whitsett JA. Function and regulation of expression of pulmonary surfactant-associated proteins. Biochem J. 1991;273(Pt 2):249–64.
Weibel ER. On the tricks alveolar epithelial cells play to make a good lung. Am J Respir Crit Care Med. 2015;191:504–13.
Whitsett JA. Genetic disorders of surfactant homeostasis. Paediatr Respir Rev. 2006;7 Suppl 1:S240–2.
Whitsett JA, Weaver TE. Hydrophobic surfactant proteins in lung function and disease. N Engl J Med. 2002;347:2141–8.
Whitsett JA, Nogee LM, Weaver TE, Horowitz AD. Human surfactant protein B: structure, function, regulation, and genetic disease. Physiol Rev. 1995;75:749–57.
Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26:1572–3.
Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics. 2015;31:1974–80.
Xu Y, Clark JC, Aronow BJ, Dey CR, Liu C, Wooldridge JL, Whitsett JA. Transcriptional adaptation to cystic fibrosis transmembrane conductance regulator deficiency. J Biol Chem. 2003;278:7674–82.
Xu Y, Liu C, Clark JC, Whitsett JA. Functional genomic responses to cystic fibrosis transmembrane conductance regulator (CFTR) and CFTR(delta508) in the lung. J Biol Chem. 2006;281:11279–91.
Xu Y, Ikegami M, Wang Y, Matsuzaki Y, Whitsett JA. Gene expression and biological processes influenced by deletion of Stat3 in pulmonary type II epithelial cells. BMC Genomics. 2007;8:455.

Xu Y, Saegusa C, Schehr A, Grant S, Whitsett JA, Ikegami M. C/EBP{alpha} is required for pulmonary cytoprotection during hyperoxia. Am J Physiol. 2009;297:L286–98.
Xu Y, Zhang M, Wang Y, Kadambi P, Dave V, Lu LJ, Whitsett JA. A systems approach to mapping transcriptional networks controlling surfactant homeostasis. BMC Genomics. 2010;11:451.
Xu J, Liu M, Xia Z. Asian medicine: call for more safety data. Nature. 2012a;482:35.
Xu Y, Wang Y, Besnard V, Ikegami M, Wert SE, Heffner C, Murray SA, Donahue LR, Whitsett JA. Transcriptional programs controlling perinatal lung maturation. PLoS One. 2012b;7:e37046.
Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013;500:593–7.
Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol. 2013;20:1131–9.
Yin H, Marshall D. Microfluidics for single cell analysis. Curr Opin Biotechnol. 2012;23:110–9.
Yin Z, Gonzales L, Kolla V, Rath N, Zhang Y, Lu MM, Kimura S, Ballard PL, Beers MF, Epstein JA, et al. Hop functions downstream of Nkx2.1 and GATA6 to mediate HDAC-dependent negative regulation of pulmonary gene expression. Am J Physiol. 2006;291:L191–9.
Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347:1138–42.
Zuo F, Kaminski N, Eugui E, Allard J, Yakhini Z, Ben-Dor A, Lollini L, Morris D, Kim Y, DeLustro B, et al. Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. Proc Natl Acad Sci U S A. 2002;99:6292–7.

Chapter 20
Functional Genomics-Renal Development and Disease

S. Steven Potter

Abstract Developmental biologists are interested in the question of how a fertilized egg, equipped with only about 22,000 genes in its nucleus, is able to transform into a complete human being. In this chapter we discuss the functional genomics analysis of mammalian kidney development, using the kidney as a model system to better understand the basic principles of organogenesis. The formation of an organ requires a complex orchestration of gene expression in many different cell types at multiple developmental stages. The driving gene expression patterns can be captured in a variety of ways. Laser capture microdissection (LCM) can be used to isolate developmental compartments of the kidney, such as the forming glomerulus, for gene expression profiling. Transgenic lines of mice can be used to fluorescently label specific cell types that can then be purified by fluorescent activated cell sorting (FACS). And more recent very high resolution technologies allow high throughput RNA-seq of single cells of a developing organ. These methodologies produce immense datasets that require powerful informatics tools for their analysis. The purpose of this chapter is to illustrate how these various tools can be used to address important questions in developmental biology that are highly relevant to child health, and development of the kidney. The huge amounts of data generated need to be captured and annotated in a systematic way, stored, integrated and analyzed. The goal is to identify basic principles as well as precise pathways that drive organogenesis. The results will provide a better understanding of developmental disorders, and guide efforts to recapitulate organogenesis in vitro, for example in the generation of replacement organs from induced pluripotent stem cells.

Keywords Development • RNA-seq • Single cell • Kidney

S.S. Potter, Ph.D. (✉)
Department of Pediatrics, Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine,
3333 Burnet Avenue, Cincinnati, OH 45229, USA
e-mail: steve.potter@cchmc.org

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7_20

20.1 Introduction

Developmental biologists are interested in the question of how a fertilized egg, equipped with only about 22,000 genes in its nucleus, is able to transform into a complete human being. In this chapter we discuss the functional genomics analysis of mammalian kidney development, using the kidney as a model system to better understand the basic principles of organogenesis.

The formation of an organ requires a complex orchestration of gene expression in many different cell types at multiple developmental stages. The driving gene expression patterns can be captured in a variety of ways. Laser capture microdissection can be used to isolate developmental compartments of the kidney, such as the forming glomerulus, for gene expression profiling. Transgenic lines of mice can be used to fluorescently label specific cell types that can then be purified by fluorescent activated cell sorting (FACS). And more recent very high resolution technologies allow high throughput RNA-seq of single cells of a developing organ. These methodologies produce immense datasets that require powerful informatics tools for their analysis. The purpose of this chapter is to illustrate how these various tools can be used to address important questions in developmental biology that are highly relevant to child health, and development of the kidney. The huge amounts of data generated need to be captured and annotated in a systematic way, stored, integrated and analyzed. The goal is to identify basic principles as well as precise pathways that drive organogenesis. The results will provide a better understanding of developmental disorders, and guide efforts to recapitulate organogenesis in vitro, for example in the generation of replacement organs from induced pluripotent stem cells.

The primary job of kidneys is to remove the end products of cellular metabolism. Kidneys also regulate blood pressure by controlling blood volume, by controlling blood salt levels, and through the production of renin, which regulates vasoconstriction through the renin-angiotensin system. The kidney is also the source of important hormones such as calcitriol, which regulates blood calcium levels, and erythropoeiten, which regulates erythropoiesis. The kidneys are only about 0.5 % of body weight, yet receive 20 % of cardiac output. They produce about 180 l of blood filtrate per day, 99 % of which is reabsorbed. The kidneys filter the entire plasma volume about 60 times per day.

The functional unit of the kidney is the nephron (Fig. 20.1). The glomerulus is the filtration element, with fenestrated capillary endothelial cells providing the blood ready access to the glomerular basement membrane filter (Fig. 20.2). The filtrate then flows through the proximal tubule, the loop of Henle, the distal tubule, and finally the collecting duct, with selective uptake and concentration occurring along the way.

The kidney provides an excellent of example of how ontogeny, the development of the individual, recapitulates phylogeny, the evolution of the species. Three kidneys, the pronephros, mesonephros and metanephros successively form during mammalian development, with each showing greater complexity, reflecting the

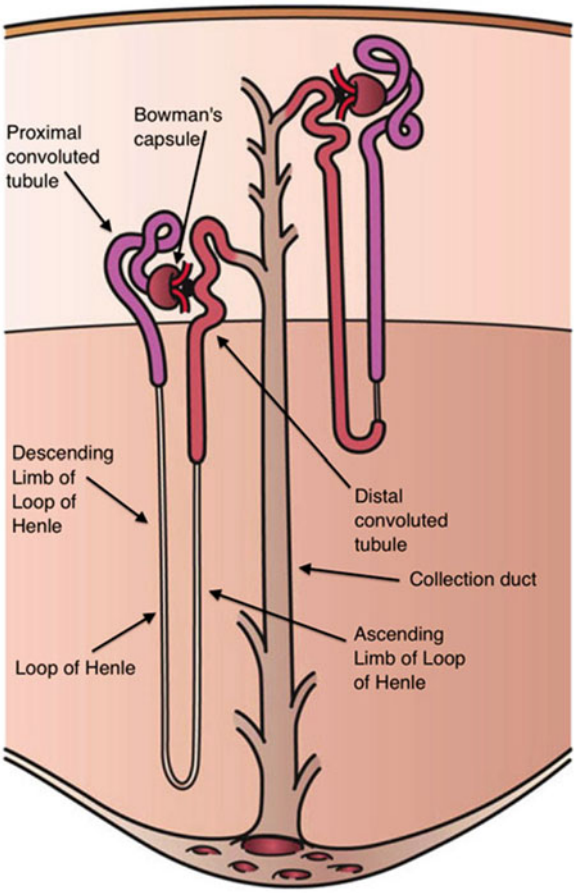


Fig. 20.1 Diagram of the nephron, the functional unit of the kidney. (Artwork by Holly Fischer)

increasing requirement for water conservation as animals moved from water to land. In this chapter we focus on the final metanephric kidney, which functions in the adult mammal.

The developing kidney provides a classic example of mutual inductive interactions (Costantini and Kopan 2010). An outgrowth of the nephric duct, the ureteric bud, invades a flanking region of condensed mesenchyme, the metanephric mesenchyme. There is essential crosstalk between the bud and mesenchyme, with the mesenchyme driving branching morphogenesis of the bud, which eventually forms the collecting duct system, and the bud inducing the mesenchyme to form nephrons. During the process of nephron formation there are multiple stages that are morphologically distinct (Fig. 20.3). The nephron progenitor cells flank the branching ureteric bud and are called capping mesenchyme. Following induction they first form a renal aggregate, which undergoes mesenchyme to epithelial transformation, making the renal vesicle, a spherical structure. Clefts then form on the renal vesicle, sequentially making the comma and S-shaped bodies, which fuse to the ureteric

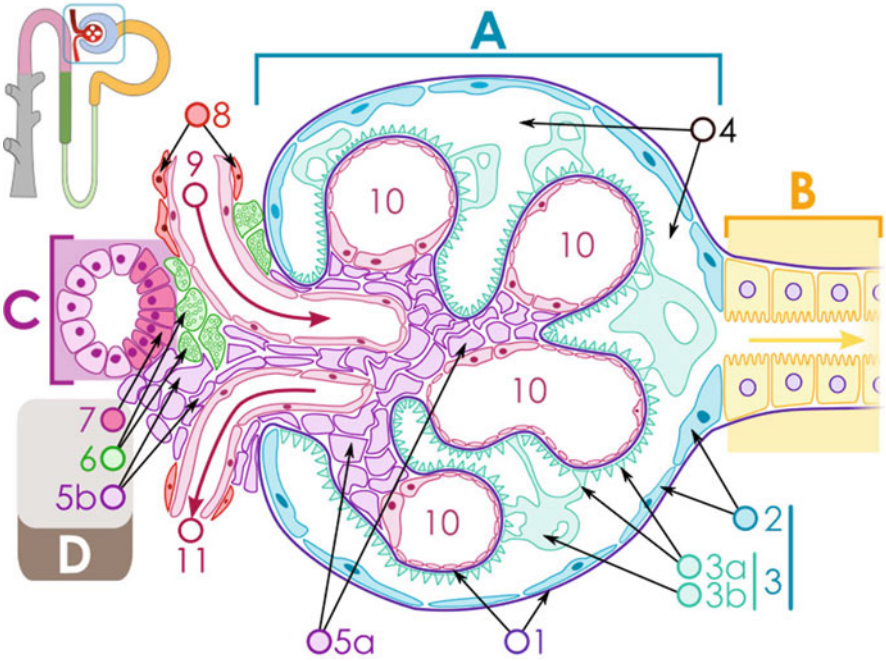


Fig. 20.2 Cross section of renal corpuscle. *A*. Renal corpuscle. *B*. Proximal tubule. *C*. Distal convoluted tubule. *D*. Juxtaglomerular apparatus. *1*. Basement membrane (Basal lamina). *2*. Bowman's capsule- parietal layer. *3*. Bowman's capsule-visceral layer. *3a*. Foot processes of podocytes. *3b*. Podocytes. *4*. Bowman's space (urinary space). *5a*. Mesangium, intraglomerular. *5b*. Mesangium, extraglomerular. *6*. Juxtaglomerular cells. *7*. Macula densa. *8*. Myocytes. *9*. Afferent arteriole. *10*. Glomerulus capillaries. *11*. Efferent arteriole (Artwork by Michal Komorniczak)

bud. The S-shaped body in turn elongates and differentiates to form the glomerulus and the various tubule segments of the mature nephron.

The mouse is a powerful model for the study of human development. All mammals have very close to the same number of genes, about 22,000. For over 16,000 human and mouse genes there is a one to one (orthologous) relationship (http://www.informatics.jax.org/homology.shtml). In addition, developmental mechanisms in general appear to be extremely well conserved during evolution. Indeed, in some cases it has been possible to demonstrate developmental functional equivalence of orthologous fruit fly and human genes through gene swap experiments. It is therefore quite likely that what we learn through the study of mice will translate well to the human system.

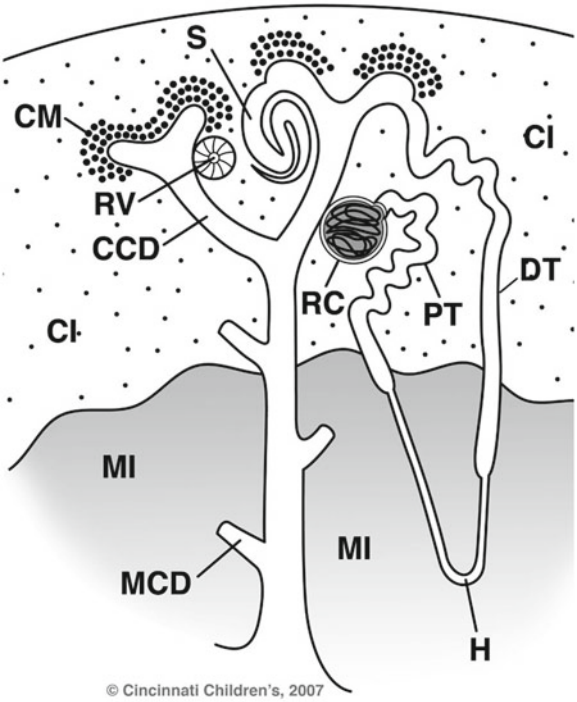


Fig. 20.3 Diagram of kidney development. *CM* capping mesenchyme progenitor cells, *RV* renal vesicle, the first epithelial precursor of the nephron, *S* S-shaped body, derived from the rrenal vesicle, *CCD* cortical collecting duct, *CI* cortical interstitium, *MI* medullary interstitium, *MCD* medullary collecting duct, *RC* renal corpuscle, *PT* proximal tubule, *H* Loop of Henle, *DT* distal tubule

20.2 Transcription Profiling and Kidney Development: Early Work

The Indian tale of the blind men and the elephant illustrates the fallacy of studies of limited scope. We can reach the wrong conclusions, or at least derive a very incomplete view, by carrying out an investigation with a focus that is too narrow. The great advance of genomics approaches is the novel ability to study, for example, all genes at once. Microarrays first, and now RNA-Seq, give global, sensitive and quantitative measures of gene expression. It is no longer necessary to make good or lucky guesses concerning which genes might be involved in a process. Indeed, it is no longer acceptable to restrict an analysis to just a few candidate genes. The challenge of the genomics era, however, is dealing with the veritable flood of data that comes from genomics tools.

Early studies used microarrays to define gene expression profiles of whole kidneys at different developmental stages (Stuart et al. 2001). These studies described for the first time the complete collection of genes expressed during nephrogenesis, and gave a measure of the changing gene expression profiles as a function of developmental time. A major limitation of these early studies, however, was a complete lack of spatial definition of gene expression. The analysis was of entire developing kidneys. Genes were identified as expressed, but the localization of the expression

remained unknown. This is the result of the homogenization of an entire organ for gene expression profiling. It was therefore difficult to define developmental pathways or cross inductive interactions of specific compartments of the forming kidney when one could not be sure which genes were being expressed in which cell type. It was clearly necessary to generate a higher resolution spatial definition of gene expression patterns in order to generate a useful genomics level understanding of kidney development.

20.3 Compartment Specific Analysis

Subsequent studies used manual microdissection as well as laser capture microdissection (LCM) to isolate the multiple compartments of the developing kidney (Brunskill et al. 2008; Challen et al. 2005; Schmidt-Ott et al. 2005; Schwab et al. 2003). It was thereby possible to purify capping mesenchyme, ureteric bud, renal vesicles, S-shaped bodies, glomeruli, proximal tubules, distal tubules, the capsule, and several different regions of stromal cells and the developing collecting duct. Gene expression profiling of these compartments resulted in the first molecular atlas of the developing kidney at microanatomic resolution.

The most comprehensive compartment specific dataset was primarily generated with LCM (Brunskill et al. 2008). The data was examined with GeneSpring, thereby defining lists of differentially expressed genes, performing clustering and producing heatmaps. The ToppGene web tool (http://toppgene.cchmc.org/), created by bioinformaticians at Cincinnati Children's Hospital, was then particularly useful in the functional analysis of the resulting extensive microarray dataset. It allows one to simply submit a list of differentially expressed genes and to receive an immediate output of enriched molecular functions, biological processes, cellular components, human phenotypes, mouse phenotypes, domains, pathways, pubmed references, interactions, cytobands, transcription factor binding sites in promoters, candidate regulatory microRNAs, and more.

Several interesting principles emerged. First, there were surprisingly few genes with a strict compartment specific expression pattern at early stages of development. For example, the capping mesenchyme and its developmental derivative, the renal vesicle, and the subsequent S-shaped bodies, all showed a considerable overlap in gene expression profile. This principle broke down, however, during the later stages of development, as compartments began to express large numbers of specific differentiation related genes. For example the proximal tubules showed relatively restricted expression of thousands of genes, including transporters involved in selective uptake.

In addition a phenomenon called anticipatory gene expression was observed. It has also been referred to as lineage priming. It appears that developmental compartments can anticipate their subsequent developmental stage and initiate expression of genes that will then peak in expression in the next compartment. For example the renal vesicle will begin to express genes that will then show much higher expression

in the subsequent S-shaped bodies. This makes sense, since the renal vesicle must start to express S-shaped body related genes as it progresses to the S-shaped body stage. Lineage priming contributes significantly to overlapping gene expression patterns.

20.4 Cell Type Specific Gene Expression Profiling

Compartment specific gene expression studies provided a considerable improvement over previous work analyzing whole homogenized organs. Nevertheless, there remained much room for improvement. A single structure, such as a developing glomerulus, will include several distinct cell types. The same logic that concludes a compartment specific atlas is superior to a gene expression profile of total kidneys also determines that a cell type specific atlas is more valuable still.

It is possible to make transgenic mice that express fluorescent reporter proteins in a cell specific manner. For example a promoter from a gene that shows cell type restricted expression can be connected to GFP (green fluorescent protein) and used to make transgenic mice. One strategy for making transgenic mice is zygote pronuclear microinjection. The DNA construct is injected into a pronucleus of a fertilized egg, which is then implanted into the oviduct of a surrogate mother. Some of the resulting offspring show stable and random integration of the DNA construct into the genome. A disadvantage of this method is that different transgenic lines, with differing chromosomal integration sites can show distinct patterns of resulting gene expression. This problem can be partially overcome by using BAC transgenics (Johansson et al. 2010). A BAC (bacterial artificial chromosome) is quite large, in some cases well over 100 kb, and can provide more regulatory sequence, giving the desired gene expression in a higher percentage of chromosomal integration sites. An alternative strategy, using a "knockin" approach, can provide more reproducible results (Mikkola and Orkin 2005). In this case a DNA construct including a reporter, such as GFP, is directly inserted into a gene showing cell type restricted expression. This is accomplished via homologous recombination in embryonic stem cells. The key advantage of this approach is that all of the regional regulatory elements of the endogenous genes are harnessed, thereby more reproducibly achieving the desired reporter expression outcome.

For example a BAC transgenic *Mafb-GFP* mouse specifically marks the podocyte in the kidney (Fig. 20.4). It is therefore possible to take the embryonic kidney, to carry out rapid enzymatic dissociation to single cells, and to use FACS to purify the GFP labeled podocytes. Similar strategies can be used to isolate other cell type specific populations of cells from the kidney. *Tie2-GFP* marks endothelial cells, and *Crym-GFP* transgenic mice allow purification of the cap mesenchyme progenitor cells.

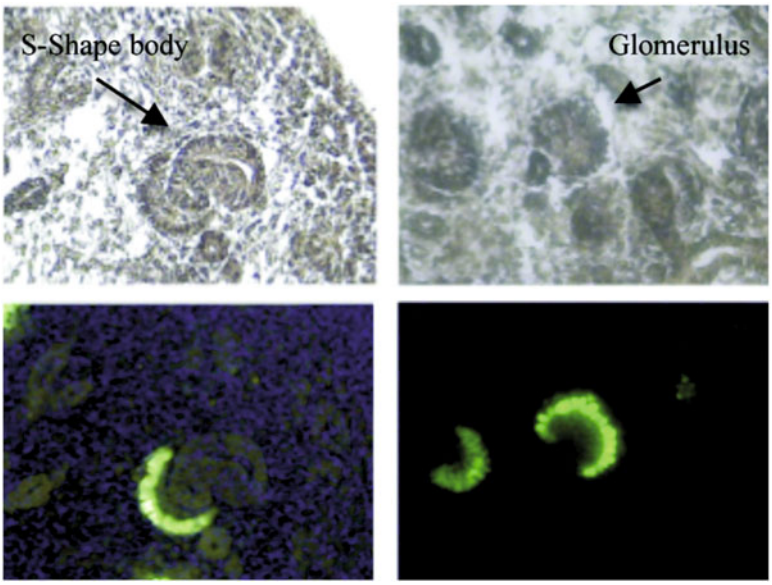


Fig. 20.4 MafB-GFP specifically marks the developing and differentiated podocytes. Brightfield images on *top* and GFP fluorescent panels on the *bottom*. Note that even early forming podocytes in the S-shaped body (*left panels*) show GFP signal

20.5 RNA-Seq

The DNA sequencing revolution is having an enormous impact on biology and medicine. The Human Genome Project took 13 years, at a total cost of about three billion dollars, to define the first human DNA sequence. But, as of this writing, it is now possible to completely sequence a human genome in a matter of days for around one thousand dollars. So, the price of the human genome has dropped over a million fold, and it continues to plummet. Over a dozen entirely new technologies are being developed, and older technologies continue to be refined.

This revolution is the foundation of "precision" or "personalized" medicine movement, based on the exact definition of the patient's genome. Different people metabolize drugs differently, and different mutations in cancers, for example, call for distinct treatments.

DNA sequencing advances are also driving a dramatic transformation in the field of gene expression profiling. It is possible to easily convert RNA to a DNA copy using the enzyme reverse transcriptase (Temin and Baltimore 1972). The complementary DNA (cDNA) copies can then be subjected to high throughput DNA sequencing, generating many millions of short reads, typically 75–100 bases in length. The reads can be aligned back to specific genes, based on their sequence. A gene that is strongly expressed will be actively transcribed to make many RNA copies. These will result in many DNA copies of the corresponding RNA, which will

result in many DNA sequence reads. Therefore by counting the number of reads corresponding to the multiple genes of the genome it is possible to generate a global digital gene expression readout.

RNA-seq provides important advantages over the older microarray technology (Mortazavi et al. 2008). First, the results are not restricted to the sequences selected for the microarray. With RNA-seq you generate sequence from any RNA present, with no bias and no selection required. Second, RNA-seq has an extremely low background. All sequences that can unambiguously be aligned with the genomic sequence, including known and novel exons, and any splice combinations. Sequences that cannot be aligned are discarded. With arrays, however, there are background and cross-hybridization issues that reduce sensitivity and can confuse results. This difference in background appears to be a very important factor in the superiority of RNA-seq over microarrays. Third, digital gene expression readouts using RNA-seq have a very wide dynamic range, estimated to cover about five orders of magnitude when using 40 million reads per experiment with the mouse genome (Mortazavi et al. 2008). Microarrays, on the other hand, have detection problems with very low abundance transcripts because of background issues, and can saturate, causing loss of linear response at high transcript levels. Thus, microarrays are limited to an effective dynamic range of only a few 100-fold. Fourth, RNA-Seq is extremely accurate in measuring transcript levels, as determined by qPCR validation (Nagalakshmi et al. 2008) and RNA spike controls (Mortazavi et al. 2008). Fifth, RNA-Seq gives an excellent view of alternative RNA processing events. It does this in two ways, by producing a digital quantitative expression level for each exon, and by providing sequence spanning exon junctions. Recent estimates are that more than 90 % of genes undergo alternative processing. In some cases the differentially processed products actually have opposite biological functions. Sixth, promoter start sites can be identified. Finally, RNA-Seq results are highly reproducible (Cloonan et al. 2008; Nagalakshmi et al. 2008).

20.6 Mouse Models of Human Kidney Disease

As an example of cell type specific gene expression let us consider the *Cd2ap* mutant mouse model of focal glomerular sclerosis (FSGS). FSGS is a progressive and devastating disease that is among the most common causes of kidney failure. The *Cd2ap* gene has a critical function in kidney podocytes, where it encodes an adaptor protein that interacts with nephrin and podocin (Schwarz et al. 2001; Shih et al. 2001). Mutations in the *Cd2ap* gene are one cause of FSGS in humans (Lowik et al. 2007). Mice with *Cd2ap* mutations also develop a severe nephrotic syndrome, including glomerulosclerosis, and die within weeks of birth (Shih et al. 2001). They therefore represent an important mouse model of this disease.

The primary deficit in FSGS appears to be in podocytes. The podocytes are quite remarkable cells (Fig. 20.2). They surround the glomerular capillaries, cooperating with the endothelial cells to synthesize the glomerular basement membrane.

Although mesodermal in origin they exhibit a striking shape, with multiple axonal like projections that reach around the capillaries. From these extend still smaller projections, the foot processes, which precisely interdigitate, leaving between them the slit diaphragms, for the passage of glomerular filtrate.

Many important functions have been associated with podocytes. The slit diaphragm is an extracellular extension of the podocyte and provides the final filtration barrier. Podocytes also function as pericytes, counteracting the hydrostatic pressure within the capillaries. In addition the podocytes are thought to play an important role in cleaning the glomerular basement membrane filter, thereby preventing clogging. Furthermore, as noted, podocytes are thought to represent initial sites of injury for a number of kidney diseases, including diabetic nephropathy and focal segmental glomerulosclerosis. Podoctye effacement and loss are among the earliest pathologic events in these diseases.

Although the podocyte is the initial site of injury in FSGS there are subsequent pathologic changes in the other two major cell types of the glomerulus, the mesangial cells and the endothelial cells. There is therefore useful knowledge to be gained through the analysis of the changing gene expression profiles in each of these cells. And each cell type, as previously noted, can be FACS purified using an appropriate transgenic-GFP mouse, with *MafB-GFP* for podocytes, *Meis1-GFP* for mesangial cells and *Tie2-GFP* for endothelial cells.

RNA-seq gene expression profiling provides a digital measure of the expression level of all genes. This is sometimes referred to as a "firehose flood" of data. In a typical experiment around 10,000 genes might show significant expression. The bioinformatics challenge is to derive important biological insight from these enormous datasets. Such insight requires extensive filtering to identify the genes of greatest interest.

An initial filter is typically the removal of genes with little or no expression in any of the samples. Genes with low expression are subject to greater variation, or noise, and can lead to artifact difference calls. Therefore these so-called "cellar dwellars" need to be filtered out. Exactly where to draw the inclusion/exclusion expression level line is subjective, but it will typically be in the range of 5–10 RPKM. That is, for example, genes with less than 10 RPKM expression level in all samples are removed from further analysis.

RPKM (or the equivalent FPKM) is a partially normalized measure of gene expression level. It stands for reads per kilobase of cDNA per million reads. It normalizes for the size of the RNA encoded by the gene, as larger complementary DNAs will give more DNA sequence reads per transcript than smaller ones. It also normalizes for the number of reads generated in the gene expression profile analysis, as more total reads will obviously give more reads per gene.

In the analysis of the *Cd2ap* mice another key filter is the comparison of the mutants to wild types. Which genes changed in expression as a result of the mutation? What gene changes are causing the disease? These are clearly the genes of greatest interest. The next stage of analysis is therefore usually a statistical test to identify differentially expressed genes. If there are just two sample types being

compared then this is an unpaired *T*-test, and for multiple types of samples, analysis of variance (ANOVA) is used, which generalizes the *T*-test to multiple groups. Filtering for P values of less than 0.05 is typical. The next consideration is choice of multiple testing correction methods. The purpose of this is to take into account the large number of data points in the analysis and to reduce the number of artifact difference calls. The Bonferroni correction, in simple terms, multiplies the uncorrected P value times the number of probesets in the analysis. For example, if 10,000 probe sets survive the expression level filter then the *T*-test P values would all be multiplied by 10,000. This is an extremely stringent correction and generally results in very short gene lists, with many truly differentially expressed genes missed. The Benjamini and Hochberg correction is somewhat less stringent. In this case the best P-value is multiplied by the total number of probe sets, as per Bonferroni, but the next best P-value is multiplied by the number of probe sets minus one, and so on. That is, as the P-values increase the correction decreases. Nevertheless, the Benjamini and Hochberg correction is still quite stringent when dealing with a large number of probesets, and again can give very short gene lists that miss many genuine differences. The Storey multiple testing correction methods are less stringent still, and are generally more appropriate unless the number of probesets has been severely trimmed prior to statistical analysis. In practice, in our experience, it is often preferred to use no multiple testing correction, and to appreciate that while this eliminates most false negatives, it results in the inclusion of some false positives. The gene list is then subjected to further filtering based on fold change, normally requiring two or threefold change minimum. This can effectively remove most of the false positives. We typically observe that over 90 % of genes that pass a screen of P value less than 0.05 and fold change greater than two can subsequently be validated as differentially expressed with an independent technology, such as quantitative PCR, immunostain, or *in situ* hybridization.

Once lists of differentially expressed genes are trimmed to a reasonable size, perhaps 20–200 genes, depending on the experiment, the next step is to carry out a functional characterization of these genes to better understand their precise biological roles. ToppGene (https://toppgene.cchmc.org) can used to further analyze the molecular processes and biological functions the genes. This web tool is very simple to use. Using the Toppfun function, the gene list is pasted in, multiple sample testing correction criteria selected, and desired features checked off. This tool includes 18 features, with perhaps the GO: Biological Process and GO: Molecular Function the most commonly used. Other functions include definition of associated mouse or human phenotypes, interactions, diseases, Pubmed lists of papers associated with the genes, and potential MicroRNA regulators of the genes in the list.

In the study of the *Cd2ap* mutant mice a number of interesting findings emerged (Brunskill and Potter 2015). The podocytes showed striking upregulation in the expression of a number of proteases and also showed a strong cell death gene expression signature. The mesangial cells showed elevated expression of pathogenic profibrotic factors as well as the angiogenesis factor Vegf. Of interest, the mesangial cells also showed expression of genes that might be considered protective, including the

fibrosis antagonist Decorin. The endothelial cells showed increased expression of leukocyte adhesion factors, likely important for their recruitment to the injured glomerulus.

20.7 RNA–Seq Analysis of Early Nephrogenesis

As an example of the power of RNA-Seq let us consider the conversion of capping mesenchymal progenitor cells (CM) to the first epithelial stage of nephron development, the renal vesicle (RV). During kidney development there is a continuous outer nephrogenic zone where the forming collecting duct undergoes branching morphogenesis and induces nephron formation. The delicate balance between progenitor renewal and differentiation determines final nephron count, which has important medical consequences (Keller et al. 2003). The induction of CM to RV, with mesenchyme to epithelia transition, is the remarkable first step in the conversion of an amorphous mesenchyme cloud into the intricate complexity of the nephron.

Many genetic regulators of early kidney development have been previously identified. WNT9b induction, as well as decreased *Six2* expression, are required for nephrogenesis (Carroll et al. 2005; Self et al. 2006). Other genes, including *Sall1*, *Fgfr1*, *Fgfr2*, *Fgf8*, *Wt1* and *Bmp7* play important roles in regulating the renewal/differentiation balance (Costantini and Kopan 2010). In addition, the LCM/microarray approach had previously identified over a 1000 genes that undergo differential expression as the CM forms the RV (Brunskill et al. 2008).

Nevertheless, despite this rather considerable prior analysis, notable new insights were derived from a further study of this process using RNA-seq (Brunskill and Potter 2012). Of interest, when the gene lists resulting from microarray and RNA-Seq were compared it was observed that over 90 % of genes called differentially expressed by microarrays were validated by RNA-seq data. We were surprised to find, however, that in addition RNA-seq found many more differentially expressed genes than microarrays, even when more stringent P-value/fold change screening criteria were used. We suspect, again, that this dramatically improved detection rate is related to the absence of background and the greater dynamic range for RNA-Seq data compared to microarray.

In this RNA-Seq study of CM differentiation into RV the data was primarily analyzed using Avadis NGS software. The workflow is very similar to that of GeneSpring microarray analysis software, allowing filtering to remove low level expressed genes, statistical testing for differential gene expression, clustering, and functional analysis of resulting gene lists.

RNA-seq can be performed to detect only polyadenylated RNAs or all RNAs, including those without polyadenylation. For example the Illumina TruSeq method first selects for polyA+ RNA and then uses random primers to generate cDNA from their entire lengths. In contrast, the Nugen Ovation RNA-Seq System V2 method has no step for selection of polyadenylated RNAs. The Nugen method avoids the generation of overwhelming amounts of cDNA from ribosomal RNA by using a

"semi random" primer that is depleted in sequences homologous to rRNAs. In the RNA-seq analysis of the CM to RV transition both of these methods were used (Brunskill and Potter 2012). This allows the two independent technologies to provide cross validation of each other, at least for polyadenylated RNAs, and it also permits a dataset comparison that determines if transcripts are polyadenylated or not.

RNA-seq data is primarily used to give a digital readout of gene expression levels and to define RNA processing patterns, but RNA-Seq data can also be used to map the positions of enhancers, many of which are transcribed (De Santa et al. 2010). For example a study of neuronal enhancers found that about half are transcribed, giving rise to short (under 2 kb) bidirectional transcripts (Kim et al. 2010). Another study of the response to macrophage activation found that PolII transcription of enhancers produced unstable transcripts (Ghisletti et al. 2010). Therefore, by performing the CM/RV RNA-Seq with both TruSeq polyA+ and Nugen random primer technologies it was possible to map the positions of "non-canonical" (not associated with standard exons) transcripts and to determine their polyadenylation status. Transcribed enhancer maps were made for both CM and RV, and it was observed that most enhancer RNAs were not polyadenylated.

Another interesting result of the CM/RV RNA-Seq study was the striking view of the Hox clusters that emerged (Brunskill and Potter 2012). First, almost all of the 39 genes of the four Hox clusters were transcribed in both CM and RV, with only the genes at the extreme ends of the clusters not expressed. For example all of the genes of the HoxA cluster were expressed in both CM and RV, except for *Hoxa1* and *Hoxa13*. Second, not only were the standard genes expressed, but the intergenic regions were also extensively transcribed, giving rise to mostly non-polyadenylated transcripts (Fig. 20.5). This widespread transcription was remarkable, and could not be attributed to repeat sequences, such as SINES and LINES, which are almost entirely absent in Hox clusters. Third, there were novel exons that were used extensively, and these often dramatically altered the function of the encoded proteins. For example in the CM a noncanonical 5′ exon was used for *Hoxd11* that gave a frameshift and as a result was noncoding, while in the RV the standard two exons were used. In addition there was extensive intergenic RNA processing and opposite strand transcription. Indeed, the picture that emerged suggested that the Hox clusters should be considered single supergenes with a multitude of transcriptional and processing possibilities rather than clusters of simple two exon genes (Brunskill and Potter 2012).

In summary, RNA-seq is revolutionizing our view of the genome. With RNA-seq we see changing RNA processing patterns, enhancer transcription, new genes, long intergenic noncoding RNA transcripts, antisense transcripts, and more. As the price of DNA sequencing continues to plummet RNA-seq will continue to improve in power, allowing more reads and a more exhaustive analysis of genome wide transcription.

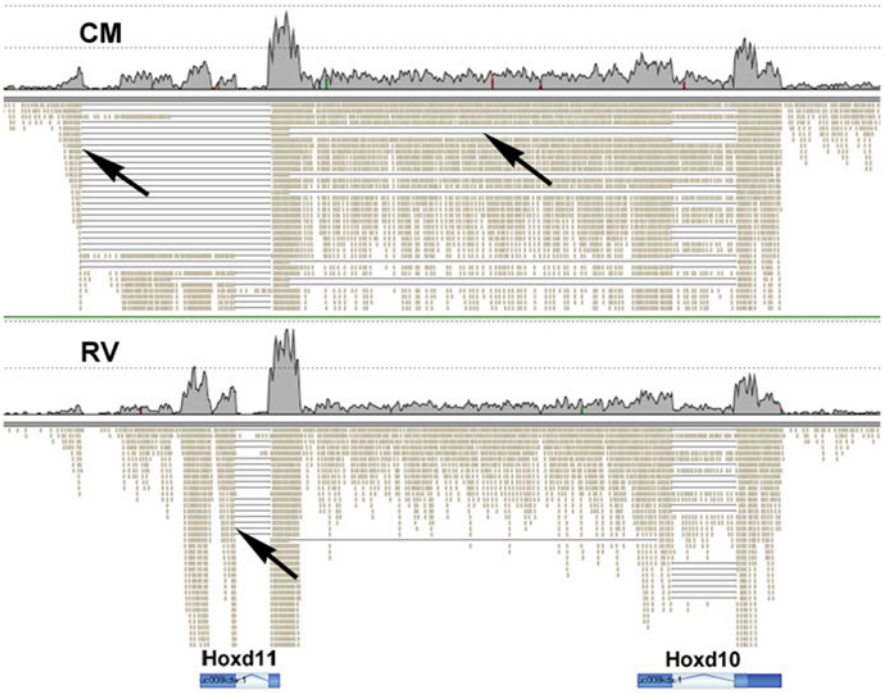


Fig. 20.5 Hox gene expression in early kidney development. Tan rectangle represent individual RNA-seq reads. Lines show introns, where the cDNA RNA-seq sequence spanned a genomic DNA region. *CM* capping mesenchyme progenitors. *RV* renal vesicle epithelial derivative of the CM. The positions of the canonical *Hoxd11* and *Hoxd10* genes are shown in *blue*. Note the extensive number of intergenic transcripts. In addition a novel 5′ exon is used for the *Hoxd11* gene in the CM, while in the RV the canonical exons are used (*arrows*). In addition there were splices that connected the two genes (*arrow*)

20.8 Single Cell Studies: Background

The RNA-Seq analysis of pure populations of cells of a specific type, such as nephron progenitors, represents a major step forward in the genomics analysis of kidney development. Nevertheless, even the thorough RNA-seq study of each specific cell type of the developing kidney would leave significant holes in our understanding. Some questions simply cannot be answered by the analysis of ensemble averages of collections of cells. They require a still higher resolution.

As mentioned earlier, development begins with the fertilized egg, which gives rise to an adult with amazing cellular variety and multiple organ systems. Underlying the enormous complexity of this process are simplifying strategies that make repeated use of individual genes and signaling systems to drive distinct developmental decisions in different contexts. At many stages in development we observe histologically uniform groups of cells that produce varied progeny. We know that

morphogen gradients, flanking cell crosstalk, and even stochastic gene expression can regulate decisions that generate diversity form apparent uniformity. Nevertheless, to date we have not created a detailed blueprint of this process, at single cell resolution, for any developing organ system.

For example, consider the early kidney progenitor cells. The metanephric mesenchyme is a histologically uniform cloud of cells that will eventually give rise to almost all of the varied cell types of the nephron. When do the first signs of differentiation into distinct developmental lineages appear within the mesenchyme? Within the early mesenchyme are some cells already lineage primed, predetermined to make specific cell types? Indeed, how varied are the gene expression profiles of the early metanephric mesenchyme? These kinds of questions can only be answered by the analysis of the transcriptional profiles of single cells.

While the genomics analysis of single cells is clearly highly desirable, there are serious challenges to consider. First there is the technical difficulty of producing an accurate gene expression profile from the extremely small quantity of RNA present in a single cell. Total RNA content per cell depends on cell type, but is generally in the range of 5–30 picograms. This is an exceedingly small amount of starting material. Some simple calculations that assume about 10 pg of total RNA per cell, with 2 % of this mRNA, determine that there are approximately 160,000 molecules of mRNA per cell. This might seem like a lot, but in fact about 10,000 genes are expressed per cell on average, so this works out to only 16 mRNA transcripts per gene. And this assumes that all mRNAs are of similar abundance, but in fact a typical cell has around a 100 or so genes that are expressed at very high levels, with 1–10,000 copies per cell, accounting in total for about half of the mass of mRNA. The end result is that for the majority of genes expressed the transcripts are present at an abundance level below ten copies per cell. And our ability to capture these RNAs, in terms of reverse transcription efficiency and amplification, is quite limited, with perhaps only 10–30 % of present RNAs actually detected. The net result is a considerable level of technical noise. Our current technology is quite imperfect in the detection and quantification of such small numbers of RNA molecules.

In addition to technical noise there is biological noise to consider. There are very few RNAs present per expressed gene and one would expect to see significant random fluctuation, even if genes were expressed in a constant steady state manner. Gene expression, however, is not steady state, but instead occurs in a pulsatile bursting mode. Early work in both bacteria and eukaryotes showed that gene expression is largely an on/off process, with gradual induction increasing the percentage of cells with expression, rather than giving incremental increase of expression in each cell (Ko et al. 1990; Novick and Weiner 1957). More recent landmark studies include the use of two copies of the same promoter in a single bacterial, or yeast cell, driving expression of two different fluorescent proteins. The results elegantly demonstrated striking fluctuation, or noise, in the expression levels of two genes with identical promoters in single cells (Elowitz et al. 2002; Ozbudak et al. 2002).

Gene expression varies from cell to cell, or within a single cell as a function of time, as a result of sporadic short bursts of active transcription (Chubb et al. 2006; Golding et al. 2005; Raj et al. 2008; Ross et al. 1994; Takasuka et al. 1998).

The causes of the pulsatile nature of gene expression are not fully defined, but one model states that transcription occurs within a limited number of transcription factories in cells (Jackson et al. 1993; Osborne et al. 2004; Wansink et al. 1993). Genes would compete for occupancy of sites within a factory where they would be highly transcribed. It has been proposed that there are relatively few of these factories, on the order of hundreds per cell, which account for the bulk of the transcription of the approximately 10,000 genes that are expressed.

There are important biological consequences to the noisy nature of gene expression. Each gene is present in only two copies per cell. And, as mentioned, each expressed gene on average has relatively few transcripts per cell. These are very small numbers, statistically speaking. And the observed burst mode of gene expression dictates that variations in transcript levels are far greater than would be predicted by a simple Poisson distribution. Indeed there are so many genes with so much variation that it makes one wonder how things ever turn out right, especially during development, when correct combinations of transcription factors are thought to drive appropriate developmental destiny decisions. One strategy for biological success is to employ genetic functional redundancy. Indeed, such redundancy is generally acknowledged to be responsible for the surprisingly mild phenotypes often observed in mice with targeted homozygous mutations of single genes.

Gene noise is probably an underappreciated cause of incomplete penetrance and variable expressivity, which can persist even on isogenic genetic backgrounds.

Remarkably, it appears that in some cases the noisy nature of gene expression has actually been harnessed to drive developmental decisions. One example is the selection of specific odorant receptors during development of the olfaction system. There are over a 1000 receptors and they are activated using a random "Monte Carlo" strategy in a mutually exclusive fashion (Tsuboi et al. 1999; Vassar et al. 1993). There is also evidence that during hematopoiesis the differentiation of stem cells could be regulated by stochastic gene expression events (Chang et al. 2008). Another elegant example is the development of photoreceptors in the *Drosophila* eye, where stochastic variations in *spineless* gene expression drive photoreceptor type differentiation decisions (Wernet et al. 2006).

Despite the technical and biological challenges it is now possible to effectively perform RNA-seq analysis of single cells. Because of noise issues it is necessary to examine many single cell replicates. Limited information is obtained from each single cell. The goal is to first generate sufficiently distinct gene expression profiles to divide single cells into categories and subtypes. The RNA-seq datasets from the multiple cells of each category are then pooled to derive an accurate gene expression definition of that cell type.

This fundamental single cell strategy is of key importance. Each individual cell represents a biological replicate. Although the gene expression data for each cell is quite imperfect, the underlying principle is to divide the cells into distinct groups based on their gene expression signatures. The data from each group is then pooled,

to average out the noise and to generate a robust gene expression profile for each cell type.

How many cells must be examined in a single cell study? A general rule of thumb is that there must be at least ten representatives of each cell type. The total number of cells required therefore depends on the degree of cell heterogeneity. If, for example, only two cell types are thought to be present, and they are in roughly equal proportion, then relatively few total cells are required to achieve the minimum of ten cells per type. On the other hand, if there is great cell heterogeneity, with many types of cells present and some being quite rare, then the total number of cells required would be very high, potentially in the thousands. In single cell studies the general rule is the more cells the better, as this provides more representatives of each cell type.

20.9 Pioneering Single Cell Studies

Early studies pioneered the strategy of dissociation of tissues into single cell suspensions followed by gene expression profiling of single cells to define novel cell types. For example Chiang and Melton carried out a single cell transcript analysis of pancreas development in 2003 (Chiang and Melton 2003). They were able to examine the developing pancreas at E10.5, when the cells are morphologically uniform. The analysis of 60 single cells allowed the developing pancreas to be divided into six subtypes based on expression of distinct markers. Of particular interest, one subset of cells showed expression of a combination of markers, including *P48*, *Nkx2.2*, and *Nkx6.1*, suggesting that these cells might be progenitors for multiple cell types.

In another study, published around the same time, a similar dissociation/single cell gene expression profiling strategy was used to examine the mammalian olfactory system (Tietjen et al. 2003). The results defined a number of genes with differential expression in olfactory sensory neurons and olfactory progenitor cells. This work included extensive validation of the general procedures used for the single cell analysis. These early studies established the general strategies that would be used for the single cell dissection of developmental mechanisms.

The limiting quantities of RNA present in single cells necessitate powerful amplification methods to generate sufficient material for microarray or RNA-seq analysis. PCR methods are most often used, but are subject to amplification bias. Methods based on in vitro transcription amplification (Van Gelder et al. 1990) offer better amplification linearity. The most recent methods, however, combine the power of PCR with the hybridization of unique molecular identifiers that completely eliminate amplification bias (see below).

20.10 High Throughput Single Cell Studies

Fluidigm offers the C1 machine that combines robotics and microfluidics to facilitate high throughput single cell analysis. The C1 receives a single cell suspension, and the cells are then randomly distributed to chambers during a capture step. Individual capture sites can then be examined with a microscope to identify those with single cells. Typically 60–80 % of capture sites show single cell occupancy. The Fluidigm C1 then carries out a series of steps, including cell lysis, reverse transcription and PCR based amplification. The products are harvested and used for RNA-seq gene expression profiling. Different IFCs are offered for the processing of cells with different sizes. The original Fluidigm C1 IFC was designed with 96 chambers, but a subsequent IFC offered 800 chambers, and it is likely that this number will increase in the future.

The advent of the Fluidigm C1 microfluidics/robotics technology greatly facilitated a number of single cell gene expression profiling studies. For example consider an analysis of the developing kidney (Brunskill et al. 2014). Single cells from the early E11.5 metanephric mesenchyme progenitors were subjected to gene expression profiling, thereby distinguishing the gene expression profiles of the cells committed to make nephrons from those that will make stoma. A surprising result was that even at this very early stage of kidney development some of the progenitors showed expression of markers of differentiated cells. For example a small fraction of the early progenitors showed robust expression of *MafB*, a marker of podocytes. This sporadic expression could be confirmed by immunostain. These early cells did not yet, however, appear to be committed to making podocytes. Many of the metanephric mesenchyme cells showed expression of one or two podocyte markers, but none showed expression of a strong podocyte signature, involving many markers. At later stages in development progenitors showed more restricted potential lineages, and expression of more genes associated with those lineages. For example the renal vesicles (RV) are the progenitors of the nephron epithelia cells. Many RV single cells showed expression of five or more podocyte specific differentiation markers, suggesting they were well on the way to becoming podocytes. Nevertheless, many of these same cells also showed expression of multiple markers of other lineages, including for example proximal and distal tubules (Fig. 20.6). In summary, it appears that early progenitors are capable of expressing a few random markers of their many potential lineages, while later progenitors will express more markers of each of their now more limited lineage choices.

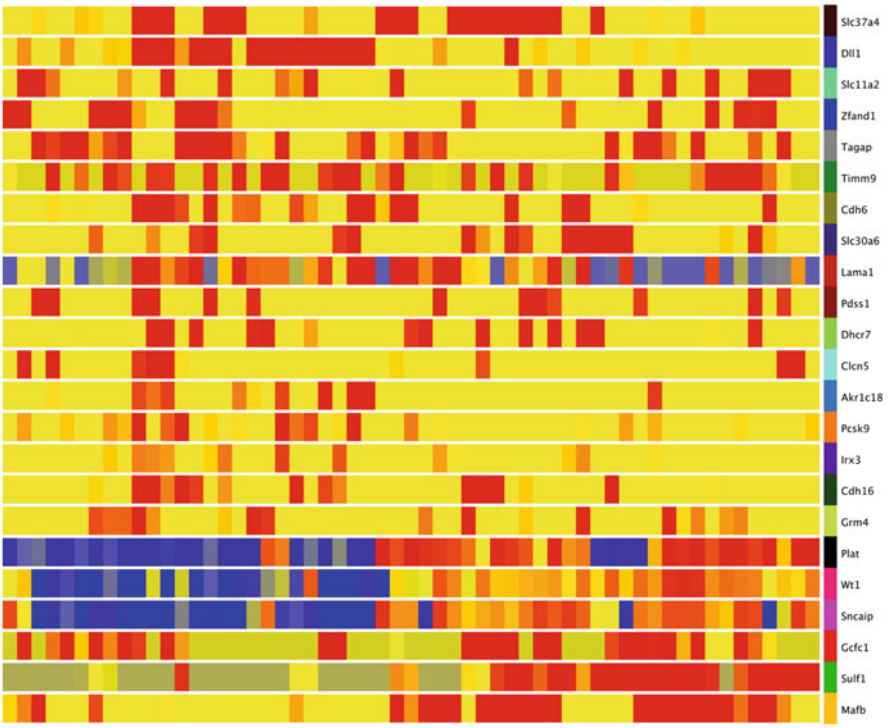


Fig. 20.6 Multilineage priming in the renal vesicle. Heatmap of renal vesicle cells, with *red* indicating high expression, *blue* is *low* expression, and *yellow* is intermediate expression. Each column is a separate single cell. Podocyte markers are Mafb, Plat, Sulf1, Sncaip and Wt1. Proximal tubule markers are Akr1c18, Clcn5, Dhcr7, Dll1, Gcfc1, Grm4, Irx3, Lama1, Pcsk9, Pdss1, Slc11a2, Slc30a6, Slc37a4, Tagap, Timm9 and Zfand1. Cdh16 is a distal tubule marker and Cdh6 is a parietal epithelial cell marker. Many cells show expression of multiple podoctye markers, even though only a small percentage of cells of the resulting nephron will be podocytes. Single cells typically show expression of multiple lineage markers

20.11 Drop-Seq, a New Single Cell RNA-Seq Technology

An exciting new technology for single cell RNA-seq was described by the McCarroll lab (Macosko et al. 2015). This Drop-seq technology is extremely high throughput, allowing the analysis of many thousands of single cells, and dramatically reduces the cost, to less than ten cents per cell, not counting the DNA sequencing costs. A simple microfluidics device is used to make aqueous drops in oil. The drops are made in a manner that results in the inclusion of a single cell in about one drop in ten, as well as a single microparticle bead. The key concept of the technique is that the beads are coated with oligonucleotides. Importantly, all of the oligonucleotides on each bead include a bead specific 12 base barcode. The cell within a drop is lysed and the polyadenylated mRNA anneals to a dT region of the oligonucleotides on the

bead. When cDNA is made from the annealed RNA it results in the inclusion of the bead specific barcode sequence. This allows all of the DNA sequence reads from the same drop to be assigned to a single cell, by virtue of the shared unique barcode.

The power of Drop-seq derives from the ability to combine all of the beads from the thousands of drops into a single tube for processing. Instead of carrying out tens of thousands of reverse transcription reactions, all are executed in a single small tube. The resulting cDNAs are cell-specific barcoded by virtue of the bead specific olidgonucleotides that they are hybridized to. The resulting remarkable efficiency is responsible for the low cost per cell for Drop-seq. The McCarroll lab used this technology to analyze the transcriptomes of 44,808 mouse retinal cells, identifying 39 distinct gene expression profile patterns, including novel cell subtypes (Macosko et al. 2015).

Another useful feature of Drop-seq is that each oligonucleotide on a bead includes an eight base Unique Molecular Identifier. So, in addition to the 12 base bead specific sequence, which is the same for all oligonucleotides on one bead, there is also an eight base sequence that is distinct for every oligonucleotide on a bead. This allows the RNA-seq reads to be aligned not only to a single cell, via the 12 base barcode, but to a specific oligonucleotide on that bead. In this manner it is possible to eliminate nonlinearity in the amplification chemistry. The RNA-seq data can therefore be deconvoluted to count the number of RNAs hybridized to the bead.

20.12 Single Cell Analysis Software Packages

Single cell RNA-seq data offers unique analysis challenges. There can be a very large number of datasets. In addition the data is noisy, so the gene expression profile generated for each cell is far from perfect. The profiles include so-called “drop-outs”, where no transcripts are detected even though the gene is actually expressed. This can result from the combination of the small number of RNAs present for the average expressed gene and the relatively inefficient capture of RNAs by the amplification chemistries. It means that cells cannot be grouped by simple single marker strategies. One cannot determine that a cell is a specific type based on the presence or absence of expression of a single gene. Instead a more complex gene expression profile is required, using the expression patterns of many genes. And these profiles are likely unknown at the start, and need to be generated as a part of the analysis.

Several software packages specifically designed to assist in the analysis of single cell data are available. A partial list includes the Singular Analysis Toolset offered by Fluidigm, Sincera (Guo et al. 2015), Monocle (Trapnell et al. 2014), AltAnalyze (Salomonis et al. 2010), and Seurat (Satija et al. 2015). Each of these programs is powerful and useful, but the complexities of single cell analysis dictate that none of them offers a simple solution. The analysis of the data clearly remains the greatest challenge in single cell gene expression profiling studies.

References

Brunskill EW, Potter SS. RNA-Seq defines novel genes, RNA processing patterns and enhancer maps for the early stages of nephrogenesis: Hox supergenes. Dev Biol. 2012;368:4–17.

Brunskill EW, Potter SS. Pathogenic pathways are activated in each major cell type of the glomerulus in the Cd2ap mutant mouse model of focal segmental glomerulosclerosis. BMC Nephrol. 2015;16:71.

Brunskill EW, Aronow BJ, Georgas K, Rumballe B, Valerius MT, Aronow J, Kaimal V, Jegga AG, Yu J, Grimmond S, et al. Atlas of gene expression in the developing kidney at microanatomic resolution. Dev Cell. 2008;15:781–91.

Brunskill EW, Park JS, Chung E, Chen F, Magella B, Potter SS. Single cell dissection of early kidney development: multilineage priming. Development. 2014;141:3093–101.

Carroll TJ, Park JS, Hayashi S, Majumdar A, McMahon AP. Wnt9b plays a central role in the regulation of mesenchymal to epithelial transitions underlying organogenesis of the mammalian urogenital system. Dev Cell. 2005;9:283–92.

Challen G, Gardiner B, Caruana G, Kostoulias X, Martinez G, Crowe M, Taylor DF, Bertram J, Little M, Grimmond SM. Temporal and spatial transcriptional programs in murine kidney development. Physiol Genomics. 2005;23:159–71.

Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature. 2008;453:544–7.

Chiang MK, Melton DA. Single-cell transcript analysis of pancreas development. Dev Cell. 2003;4:383–93.

Chubb JR, Trcek T, Shenoy SM, Singer RH. Transcriptional pulsing of a developmental gene. Curr Biol. 2006;16:1018–25.

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods. 2008;5:613–9.

Costantini F, Kopan R. Patterning a complex organ: branching morphogenesis and nephron segmentation in kidney development. Dev Cell. 2010;18:698–712.

De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. 2010;8:e1000384.

Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002;297:1183–6.

Ghisletti S, Barozzi I, Mietton F, Polletti S, De Santa F, Venturini E, Gregory L, Lonie L, Chew A, Wei CL, et al. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. Immunity. 2010;32:317–28.

Golding I, Paulsson J, Zawilski SM, Cox EC. Real-time kinetics of gene activity in individual bacteria. Cell. 2005;123:1025–36.

Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. PLoS Comput Biol. 2015;11:e1004575.

Jackson DA, Hassan AB, Errington RJ, Cook PR. Visualization of focal sites of transcription within human nuclei. EMBO J. 1993;12:1059–65.

Johansson T, Broll I, Frenz T, Hemmers S, Becher B, Zeilhofer HU, Buch T. Building a zoo of mice for genetic analyses: a comprehensive protocol for the rapid generation of BAC transgenic mice. Genesis. 2010;48:264–80.

Keller G, Zimmer G, Mall G, Ritz E, Amann K. Nephron number in patients with primary hypertension. N Engl J Med. 2003;348:101–8.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465:182–7.

Ko MS, Nakauchi H, Takahashi N. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. EMBO J. 1990;9:2835–42.

Lowik MM, Groenen PJ, Pronk I, Lilien MR, Goldschmeding R, Dijkman HB, Levtchenko EN, Monnens LA, van den Heuvel LP. Focal segmental glomerulosclerosis in a patient homozygous for a CD2AP mutation. Kidney Int. 2007;72:1198–203.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14.

Mikkola HK, Orkin SH. Gene targeting and transgenic strategies for the analysis of hematopoietic development in the mouse. Methods Mol Med. 2005;105:3–22.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320:1344–9.

Novick A, Weiner M. Enzyme induction as an all-or-none phenomenon. Proc Natl Acad Sci U S A. 1957;43:553–66.

Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell JA, Lopes S, Reik W, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet. 2004;36:1065–71.

Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A. Regulation of noise in the expression of a single gene. Nat Genet. 2002;31:69–73.

Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods. 2008;5:877–9.

Ross IL, Browne CM, Hume DA. Transcription of individual genes in eukaryotic cells occurs randomly and infrequently. Immunol Cell Biol. 1994;72:177–85.

Salomonis N, Schlieve CR, Pereira L, Wahlquist C, Colas A, Zambon AC, Vranizan K, Spindler MJ, Pico AR, Cline MS, et al. Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation. Proc Natl Acad Sci U S A. 2010;107:10514–9.

Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33:495–502.

Schmidt-Ott KM, Yang J, Chen X, Wang H, Paragas N, Mori K, Li JY, Lu B, Costantini F, Schiffer M, et al. Novel regulators of kidney development from the tips of the ureteric bud. J Am Soc Nephrol. 2005;16:1993–2002.

Schwab K, Patterson LT, Aronow BJ, Luckas R, Liang HC, Potter SS. A catalogue of gene expression in the developing kidney. Kidney Int. 2003;64:1588–604.

Schwarz K, Simons M, Reiser J, Saleem MA, Faul C, Kriz W, Shaw AS, Holzman LB, Mundel P. Podocin, a raft-associated component of the glomerular slit diaphragm, interacts with CD2AP and nephrin. J Clin Invest. 2001;108:1621–9.

Self M, Lagutin OV, Bowling B, Hendrix J, Cai Y, Dressler GR, Oliver G. Six2 is required for suppression of nephrogenesis and progenitor renewal in the developing kidney. EMBO J. 2006;25:5214–28.

Shih NY, Li J, Cotran R, Mundel P, Miner JH, Shaw AS. CD2AP localizes to the slit diaphragm and binds to nephrin via a novel C-terminal domain. Am J Pathol. 2001;159:2303–8.

Stuart RO, Bush KT, Nigam SK. Changes in global gene expression patterns during development and maturation of the rat kidney. Proc Natl Acad Sci U S A. 2001;98:5649–54.

Takasuka N, White MR, Wood CD, Robertson WR, Davis JR. Dynamic changes in prolactin promoter activation in individual living lactotrophic cells. Endocrinology. 1998;139:1361–8.

Temin HM, Baltimore D. RNA-directed DNA synthesis and RNA tumor viruses. Adv Virus Res. 1972;17:129–86.

Tietjen I, Rihel JM, Cao Y, Koentges G, Zakhary L, Dulac C. Single-cell transcriptional analysis of neuronal progenitors. Neuron. 2003;38:161–75.

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32:381–6.

Tsuboi A, Yoshihara S, Yamazaki N, Kasai H, Asai-Tsuboi H, Komatsu M, Serizawa S, Ishii T, Matsuda Y, Nagawa F, et al. Olfactory neurons expressing closely linked and homologous odorant receptor genes tend to project their axons to neighboring glomeruli on the olfactory bulb. J Neurosci. 1999;19:8409–18.

Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci U S A. 1990;87:1663–7.

Vassar R, Ngai J, Axel R. Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. Cell. 1993;74:309–18.

Wansink DG, Schul W, van der Kraan I, van Steensel B, van Driel R, de Jong L. Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. J Cell Biol. 1993;122:283–93.

Wernet MF, Mazzoni EO, Celik A, Duncan DM, Duncan I, Desplan C. Stochastic spineless expression creates the retinal mosaic for colour vision. Nature. 2006;440:174–80.

Index

J.J. Hutton (ed.), *Pediatric Biomedical Informatics*, Translational Bioinformatics 10,
DOI 10.1007/978-981-10-1104-7