

Chapter 10

Data Science: Transformation of Research and Scholarship

Lynda R. Hardy and Philip E. Bourne

Abstract The emergence of data science as a practice and discipline is revolutionizing research potential in all disciplines, but healthcare science has the potential to affect the health of individual lives. The use of existing data provides fertile ground for healthcare professionals to conduct research that will maximize quality outcomes, develop algorithms of care to increase efficiency and safety, and create predictive models that have the ability to prevent illness events and reduce healthcare expenditures. Data science can change practice through using existing and growing amounts of data to conduct research and build scholarship. Clinical trials, in some cases, may no longer be required to examine interventions. The pragmatic and efficient use of existing large cohort datasets has the ability to generate sample and control groups to determine efficacy. The collection of data from electronic medical records can provide substantial data to determine trends, construct algorithms, and consider disease and health behaviors modeling that alter patient care. *Digital research* incorporating vast amounts of data and new analytics has the ability to influence global healthcare.

Keywords Big data • Data science • Translational research • Schema-on-write
Schema-on-read

L.R. Hardy, Ph.D., R.N., F.A.A.N. (✉)
College of Nursing, The Ohio State University, Columbus, OH, USA
e-mail: unclyn@gmail.com; hardy.305@osu.edu

P.E. Bourne, Ph.D.
Data Science Institute, University of Virginia, Charlottesville, VA, USA

10.1 Introduction to Nursing Research

The purpose of this chapter is to understand how research in the digital age and big data are transforming health-related research and scholarship suggesting a paradigm shift and new epistemology.

Health science and the translation of research findings are not new. Florence Nightingale, a social reformer and statistician who laid the foundation for nursing education, conducted early health-related research during the Crimean War by collecting data on causality of death in soldiers. Nightingale was a data collector, statistician, and concerned with data visualization as indicated by her Rose Diagram, a topic of much research. The diagram originally published in *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army. Founded Chiefly on the Experience of the Late War. Presented by Request to the Secretary of State for War* graphically presented data indicating that more soldiers died because of disease than of their battle-related injuries. Nightingale was a pioneer at discovery and data-based rationale underpinning the practice of nursing and health-related implications. Informed by philosophy, she was a systematic thinker, who understood the need for systematic data collection (Fig. 10.1). Nightingale designed survey instruments and determined their validity through vetting with experts in the field. She was a statistician basing her findings on mathematics. Through using the findings of her work to change practice and policy, Nightingale reformed conditions for workhouse poor, patient care standards and the right to a meaningful death (McDonald 2001, Selanders and Crane 2012). Imagine what she might have done in the digital age, with computers, and big data.

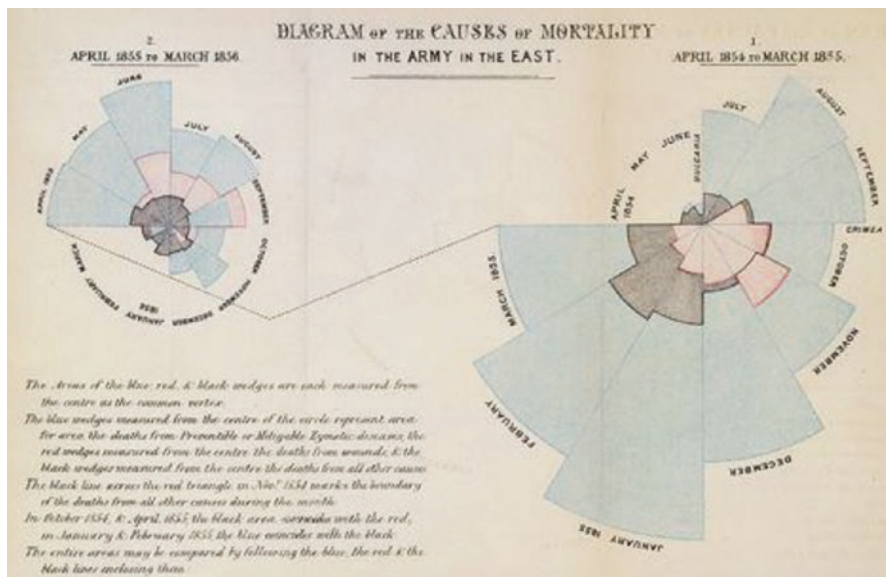


Fig. 10.1 Nightingale rose diagram. Source History of Information.com. (Accessed 5 Dec 2015)

Nursing research, as defined by the National Institute of Nursing Research, is knowledge development to “build the scientific foundation for clinical practice; prevent disease and disability; manage and eliminate symptoms caused by illness, and enhance end-of-life and palliative care” (NINR 2015). Grady and Gough (2015) further suggest that nursing science provides a bridge from basic to translational research via an interdisciplinary or team science approach to increasing the understanding of prevention and care of individuals and families through personalized approaches across the lifespan (Grady and Gough 2015).

10.1.1 *Big Data and Nursing*

The profession of nursing has been intricately involved with healthcare data since the beginning of nurse’s notes that documented patient care and outcomes. Notes and plans of care are reviewed, shared and modified for future care and better outcomes. The magnitude of healthcare data is complex thereby requiring nontraditional methods of analysis. The interweaving of multiple data streams to visualize, analyze and understand the entirety of human health demands a powerful method of data management, association, and aggregation. The digitization of healthcare data has consequently increased the ability to aggregate and analyze these data focusing on historical, current and predictive possibilities for health improvement.

The foundation of nursing research is the integration of a hypothesis-driven research question supported by an appropriate theoretical framework. The fourth paradigm, a phrase originally coined by Jim Gray (Hey et al. 2009), or data science, is considered by some to be the end of theory-based research, thus creating a reborn empiricism whereby knowledge is derived from sense experience. This view is not without its detractors. There were big data skeptics before the explosion took place. Dr. Melvin Kranzberg, professor of technology history, in his 1986 Presidential Address, commented that “Technology is neither good nor bad: nor is it neutral ... technology’s interaction with the social ecology is such that technical developments frequently have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves” (Kranzberg 1986, p. 545). Kitchin (2014) further suggested that “big data is a representation and a sample shaped by technology and platform used, the data ontology employed and the regulatory environment” (p. 4). Kitchin’s statement reinforces the idea that data cannot explain itself but requires a lens (e.g., theoretical framework) through which to interpret the data. Data is *raw/without interpretation* and cannot interpret outliers or aberrancies suggesting bias. It provides a bulk of information where a specific analytic process is applied—but at the end, the data must be interpreted. Use of a theoretical framework provides a pathway of understanding that can support statistical findings of data analysis.

Bell (2009) suggests that “data comes in all scales and sizes ... data science consists of three basic activities, capture, curation, and analysis” (p. xiii). He also comments on Jim Gray’s proposal that scientific inquiry is based on four paradigms:

Table 10.1 Research paradigms

Paradigm	Gray	General research	Research type
1	Experimental science	Positivism	Quantitative
2	Theoretical science	Anti-positivism	Qualitative
3	Computational science	Critical theory	Critical and action oriented
4	Data science		

experimental, theoretical, computational, and data science (Hey et al. 2009). Table 10.1 provides an integration of Gray's paradigms with general research paradigms. Gray's fourth paradigm supports the integration of the first three

The explosion of big data (defined by volume, velocity and veracity) in healthcare provides opportunities and challenges. Healthcare providers, researchers and academics have the ability to visualize individual participant data from multiple sources (hospital, clinic, urgent care and school settings, claims data, research data) and in many forms (laboratory, imaging, provider notes, pharmacy, and demographics). The challenge of aggregating and analyzing these data streams is possessing usable standardized data and the right analytic tools with the power to aggregate and understand data types. The outcome of having appropriate data and tools is the ability to improve health care outcomes and develop predictive models for prevention and management of illness. Other advantages include those related to clinical operations, research, public health, evidence-based medicine/care, genomic analytics, device (wearable and static) monitoring, patient awareness, and fraud analysis (Manyika et al. 2011). Platforms have been developed to assist with the analyses of the various data streams, e.g., Hadoop, Cloudera CDH, Hortonworks, Microsoft HDInsight, IBM Big Data Platform, and Pivotal Big Data Suite. These platforms frequently use cloud computing—ubiquitous elastic compute and large data storage engines from the likes of Google, Amazon and Microsoft. Thus not only is the scientific paradigm changing, but also the compute paradigm from local processing to distributed processing. These changes are accompanied by a new software industry focusing on such areas as data compression, integration, visualization, provenance and more.

More specifically analytical methods, other than using a theoretical pathway to determine what data should be collected and the method of analysis should be considered to allow the data and not the theory to provide the pathway. Many methods are becoming available as a means to analyze and visualize big data. One method, point cloud, uses a set of data points in a three dimensional system for data visualization (Brennan et al. 2015). Other methods use various forms of data clustering such as cluster analysis (groups a set of objects/data points into similar clusters) (Eisen et al. 1998) and progeny clustering (applies cluster analysis determining the optimal number of clusters required for analysis) (Hu et al. 2015). A variety of methods exist to examine and analyze healthcare data providing rich data for improving patient outcomes, predictive modeling and publishing the results.

Theory-based research uses the method of *schema-on-write* (Deutsch 2013) which is usually a clean and consistent dataset but the dataset is more limited or narrow. This method, where data were pre-applied to a plan, requires less work

initially but also may result in a more limited result. Research opportunities today provide the ability to broaden the scope and magnitude of the data by allowing for the expansion and use of multiple types/streams of data using the method of *schema-on-read*. *Schema-on-read* identifies pathways and themes at the end of the process, allows the researcher to cast a wide data net incorporating many types of structured and unstructured data, and finally applies the theoretical pathway to allow the analysis to ‘make sense’ of the data. The data is generally not standardized or well organized but becomes more organized as it is used. The data has the ability to be more flexible thus providing more information (Pasqua 2014). In summary, *schema-on-read* provides the ability to create a dataset with a multidimensional view; these traits magnifying the usability of the dataset. This expands the nurse researcher’s ability to explain the research question leading to development of preventive or treatment interventions.

The digital environment and diversity of data have created the need for interdisciplinary collaboration using scientific inquiry and employing a team science approach. This approach provides an environment to maximize self-management of illness, increase, maintain a level of individual independence, and predict usefulness of interventions within and external to professional health environments. Individual empowerment allows individuals to participate, compare outcomes, and analyze their own data. This is accomplished using simple, smart-phone applications and wearable devices. It is the epitome of self-management and participatory research.

10.1.2 *Nursing and Data*

The Health Information Technology for Economic and Clinical Health (HITECH) Act, an initiative passed in 2009, provides financial support for electronic health records (EHRs) to promote meaningful use in medical records thereby expanding EHR use and healthcare information. Nurse scientists previously gathered data from smaller data sets that were more narrowly focused such as individual small research studies and access to data points within the EHR. This approach provided a limited or constricted view of health-related issues. The unstructured nature of the data made data extraction difficult due to non-standardization. Using the narrow focus of the data inhibited the ability to generalize findings to a larger population often resulting in the need for additional studies. Moreover, EHR data focused on physician diagnosis and related data, failing to capture the unstructured but more descriptive data, e.g., nursing data (Wang and Krishnan 2014).

The big data tsunami allows nurse scientists access to multiple data streams and thus expands insight into EHRs, environmental data that provides an exposome or human environment approach, genetic and genomic data allowing for individuality of treatment and technology driven data such as wearable technology and biosensors allowing nearly real-time physiologic data analysis. Moreover, data sharing provides power that has not previously existed to detect differences in health dis-

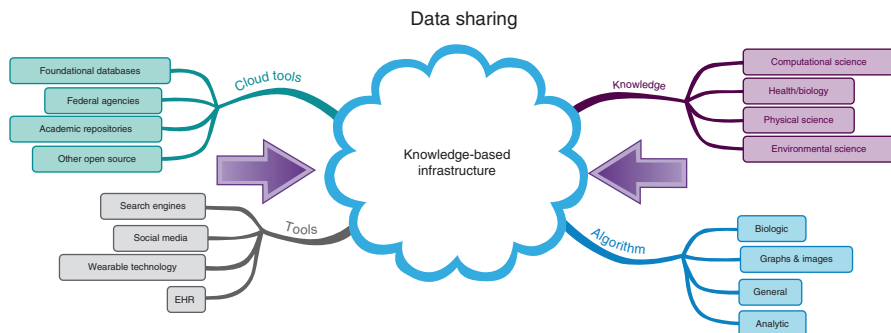


Fig. 10.2 Challenges of big data. Adapted from the high-level research design in the fourth discipline (p. 169)

parities. Data aggregation generates a large participant pool for use in pragmatic studies to understand the depth and breadth of health disparities (Fig. 10.2).

For example, technology-driven data that impacts elder care and can prevent adult injury is fall-related data collected by retrieving data from hospital or home systems (Rantz et al. 2014). This data widen the scope of evidence-based healthcare by providing a multidimensional interpretation of the data. Larger, more inclusive datasets such as claims data from the Centers for Medicare and Medicaid Services (CMS) add to the ability to create a more complete view of health and healthcare. However, this data is constrained by what claims form information is captured and available for research purposes. Data digitization and open access publication provide a rich environment for all disciplines to access, correlate, analyze and predict healthcare outcomes. Future data sharing and reuse will likely capture more of the research continuum and process.

10.2 The New World of Data Science

Data science and resulting data use are not new and are growing at warp speed. It is estimated that 2.7 zettabytes of data were generated daily in 2012; 35 zettabytes of data per day are anticipated by 2020; 90% of current data were collected in the past 2 years and 5 exabytes are generated daily (Karr 2012). A brief history of data science shows that Tukey (1962) in describing his transformation from an interest in statistics to one in data analysis initiated the thought that there was a difference between the two. It was not until the mid-1990s that a more formalized approach to data (analysis) science was developed that began to look at the increased accumulation of data and data analytics (Tukey 1962).

The advent of data science is compared to Fordism. The world changed when Henry Ford discovered new methods of building automobiles. Manufacturing processes were modernized, modifying knowledge and altering methods of

understanding the world. Fordism changed society and behaviors impacting everyone (Baca 2004). The big data explosion follows the same trajectory. New methods of capturing, storing and analyzing data have and will continue to have an impact on society. Today's data has exploded into multiple data streams, structured and unstructured.

Rapid growth of the big data or data science ecosystem emphasizes the need for interdisciplinary approaches to interpreting and understanding health and health-care data. The data explosion provides an environment with various data types require the same breadth of scientists to interpret the data accurately. Just as generation of data is from a plurality of sources so must the composition of the team assigned to its analysis. The data explosion provides nurse scientists, as well as other disciplines, the ability to work within highly diverse teams to provide deep knowledge integration and comprehensive analysis of the data. This expansive inclusion of expertise extends to employing the skills of citizen scientists.

The data explosion also provides a new world data alchemy allowing for transformation, creation and combination of data types to benefit healthcare outcomes through accurate decision-making and predictability. Data standards are becoming more prevalent. One key example of this prevalence is the NIH's Big Data to Knowledge (BD2K) program recently establishing a Standards Coordinating Center (SCC). One example of standards work is Westra and Delaney success in having the Nursing Management Minimum Data Set (NMMDS) incorporated into Logical Observation Identifiers Names and Codes (LOINC) a universal system for test, observations and measurement (Westra et al. 2008). Computers are becoming ever more powerful and due to the ubiquitous nature of data and data-driven algorithms allow deeper and more complex analyses resulting in greater accuracy in patient care-related decision making (Provost and Fawcett 2013). Making use of existing and future data necessitates training of data scientists. The need for data scientists is growing with an anticipated shortage of between 140,000 and 150,000 people (Violino 2014). The combination of these elements—standards, data, analytics and a trained workforce—increases the accuracy and predictive use of data with numerous opportunities for scholarship, including publication and analytic developments.

10.3 The Impact of Data Proliferation on Scholarship

Scholarship (FreeDictionary: academic achievement; erudition; Oxford Dictionary: learning and academic study or achievement) was once solely a paper journal publication (p-journal) and an academic requirement for tenure. The era of big data and data science increases the realm of scholarship by adding a variety of publication/dissemination forms such as electronic journal publication (e-journal), web-based formal or informal documentation, reference data sets, and analytics in the form of software and database resources. Digitalization of online information and data provide fertile ground for new scholarship. The technological provision of shared data,

cloud computing and dissemination of publications places scholarship in the fast lane for nursing and other disciplines. The information superhighway is clearly the next generation infrastructure for scholarship even as the academic establishment's adoption of such change is behind that pace of change. Such a gap and the migration of a skilled workforce of data scientists from academic research to the private sector are concerns.

“Scholarship represents invaluable intellectual capital, but the value of that capital lies in its effective dissemination to present and future audiences.”
AAU, ARL, CNI (2009)

Much of academic scholarship is based on Boyer's model espousing that original research is centered on discovery, teaching, knowledge and integration (Boyer 1990). The American Association of Colleges of Nursing (AACN) adopted Boyer's model defining nursing research as: "... those activities that systematically advance the teaching, research, and practice of nursing through rigorous inquiry that (1) is significant to the profession, (2) is creative, (3) can be documented, (4) can be replicated or elaborated, and (5) can be peer-reviewed through various methods" (2006). A hallmark of scholarship dissemination continues to be a process of peer-reviewed publications in refereed journals. The Association of American Universities (AAU), the Association of Research Libraries (ARL) and the Coalition for Network Information (CNI) published a report emphasizing the need to disseminate scholarship. Big data now questions if the scholarship model requires updating to be more inclusive of the sea change in information (culturally, socially, and philosophically) technology has introduced (Boyd and Crawford 2012).

Today's digital environment and the need for dissemination of scholarly work suggest expanding the definition to allow for other methods of scholarship. Borgman and colleagues, in their 1996 report to the National Science Foundation (NSF), developed the information life cycle model as a description of activities in creating, searching and using information (Borgman et al. 1996). The outer ring denotes life cycle stages (active, semi-active, inactive) with four phases (creation, social context, searching and utilization), where creation is the most active. The model includes six stages, which further assist the context of scholarship utilization (Fig. 10.3).

The incorporation of the information life cycle model into the AACN scholarship definition adds the need for dissemination and the inclusion of sources outside the normal process of p-journal publications. This incorporation would highlight that publication is a multi-dimensional continuum requiring three main criteria. First, the information must be publicly available via sources such as subscriptions, abstracts and databases/datasets allowing for awareness and accessibility of the work. Second, the scholarly work should be trustworthy; this is generally conducted through peer review and identified institutional affiliation. Finally, dissemination and accessibility are the third criterion that allows visualization of the scholarly work by others (Kling 2004).

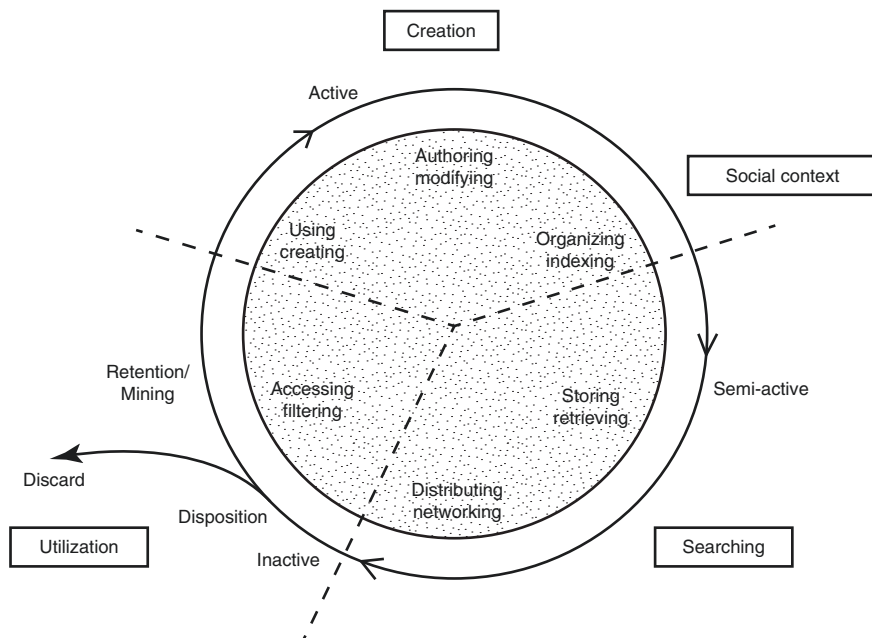


Fig. 10.3 Information life cycle. *Source* Borgman et al. (1996) report to NSF. *Note* The outer ring indicates the life cycle stages (active, semi-active, and inactive) for a given type of information artifact (such as business records, artworks, documents, or scientific data). The stages are superimposed on six types of information uses or processes (shaded circle). The cycle has three major phases: information creation, searching, and utilization. The alignment of the cycle stages with the steps of information handling and process phases may vary according to the particular social or institutional context

The digital enterprise is no longer relegated to p-journals but has increased to include e-journals, data sets, repositories (created to collect, annotate, curate and store data) and publicly shared scholarly presentations. Citations of data sets now receive credibility and validity through this new scholarship type, in part through the work of the National Institutes of Health Big Data to Knowledge (BD2K) initiative. The rapid ability to access, analyze either through cloud computing or novel software designed for big data, and disseminate through multiple methods provides nursing and all disciplines a more rapid ability to publicize and legitimize their scholarly work.

10.4 Initiatives Supporting Data Science and Research

Many government agencies have initiated work designed to build processes within the digital ecosystem to assist teams focusing on data science. These initiatives have been developed as a means of assisting in faculty and student training, collaboration with centers of excellence, development of software designed to facilitate the

analysis of large datasets, and the ability to share data and information through a cloud-based ecosystem that maximizes the use of existing multi-dimensional data to better understand and predict better patient outcomes. Further, multiple agencies have open funding sources consistent with nursing science. Examples follow.

10.4.1 National Institutes of Health

The National Institutes of Health (NIH) spearheaded the big data program with the creation of the Big Data to Knowledge or BD2K initiative in 2012 when an advisory committee convened by Dr. Francis Collins, the NIH Director, investigated the depth and breadth of big data potential. Dr. Collins and key members of his leadership team reviewed the committee's findings and committed to providing a 'data czar' to facilitate data science that would span the 27 Institutes that comprised NIH. The BD2K team, led by Dr. Phil Bourne (co-author of this work), created a data science ecosystem incorporating (1) **training** for all levels of data scientists, (2) centers that would work independently and in concert with all BD2K **centers** to build a knowledge base, (3) a **software** development team focused on creating and subsequently maintaining new methods for big data analysis, and (4) a **data indexing** team focused on creating methods for indexing and referencing datasets. (<https://datascience.nih.gov>). Taken together the intent is to make data FAIR—Findable, Accessible (aka usable), Interoperable and Reusable.

10.4.1.1 Training

Training focused on establishing an effective and diverse biomedical data science workforce using multiple methods across educational and career levels—students through scientists. Training focused on the continuum of scientists who see biomedical data science as their primary occupation to those that see biomedical data science as a supplement to their skill set. Development and funding of a training coordination center that ensures all NIH training materials are discoverable is paramount.

10.4.1.2 Centers

Centers included the establishment of 11 Centers of Excellence for Big Data Computing and two Centers that are collaborative projects with the NIH Common Fund LINCS program (the LINCS-BD2K Perturbation Data Coordination and Integration Center, and the Broad Institute LINCS Center for Transcriptomics). Centers are located throughout the United States providing training to advance big data science in the context of biomedical research across a variety of domains and datatypes.

10.4.1.3 Software

Software focus included targeted Software Development awards to fund software and methods for the development of tools addressing data management, transformation, and analysis challenges in areas of high need to the biomedical research community.

10.4.1.4 Commons

Commons addressed the development of a scalable, cost effective electronic infrastructure simplifying, locating, accessing and sharing of digital research objects such as data, software, metadata and workflows in accordance with the FAIR principles (<https://www.force11.org/group/fairgroup/fairprinciples>).

10.4.1.5 Data Index

Data Index is a data discovery index (DDI) prototype (<https://biocaddie.org/>) that indexes data that are stored elsewhere. The DDI will increasingly play an important role in promoting data integration through the adoption of content standards and alignment to common data elements and high-level schema.

10.4.2 *National Science Foundation*

The National Science Foundation (NSF) is a United States government agency supporting research and education in non-medical fields of science and engineering. NSF's mission is "to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense..." (<http://www.nsf.gov> accessed 12/23/2015). The annual NSF budget of \$7.3 billion (FY 2015) is the funding source for approximately 24% of all federally supported basic research conducted by U.S. colleges and universities.

NSF created the Directorate for Computer & Information Science & Engineering (CISE) with four goals related to data science to:

- Uphold the U.S. position of world leadership in computing, communications, and information science and engineering;
- Promote advanced computing, communications and information systems understanding;
- Support and provide advanced cyberinfrastructure for the acceleration of discovery and innovation across all disciplines; and
- Contribute to universal, transparent and affordable participation in an information-based society.

CISE consists of four divisions, each organized into smaller programs, responsible for managing research and education. These four divisions (the Division of Advanced Cyberinfrastructure; the Division of Computing & Communication Foundations; the Division of Computer and Network Systems; and the Division of Information and Intelligent Systems) incorporate program directors acting as the point of contact for sub-disciplines that work across each division and between divisions and directorates. NSF CISE provides funding in the areas of research infrastructure, advancing women in academic science and engineering, cybersecurity, big data hub and spoke designs to advance big data applications, and computational and data science solicitations to enable science and engineering (<http://www.nsf.gov/cise/about.jsp> accessed 12/23/2015).

10.4.3 U.S. Department of Energy

The U.S. Department of Energy's (DOE) mission is to "ensure America's security and prosperity by addressing its energy, environmental and nuclear challenges through transformative science and technology solutions" (<http://energy.gov/mission> accessed 12/23/2015). A prime focus of the DOE is to understand open energy data through the use and access to solar technologies. The DOE collaborated with its National Laboratories to harness data to analyze new information from these large data sets (Pacific Northwest National Lab), train researchers to think about big data, and to focus on issues of health-related data (Oak Ridge National Lab). As an example, the Oak Ridge National Laboratory's Health Data Sciences Institute, in concert with the National Library of Medicine (NLM), developed a new and more rapid process to accelerate medical research and discovery. The process, Oak Ridge Graph Analytics for Medical Innovation (ORiGAMI), is an advanced tool for literature-based discovery.

10.4.4 U.S. Department of Defense

The U.S. Department of Defense (DOD) focuses on cyberspace to enable military, intelligence, business operations and personnel management and movement. The DOD focuses on protection from cyber vulnerability that could undermine U.S. governmental security. The four DOD foci include (1) resilient cyber defense, (2) transformation of cyber defense operations, (3) enhanced cyber situational awareness, and (4) survivability from sophisticated cyber-attacks (<http://www.defense.gov> accessed 12/23/2015).

The Defense Advanced Research Projects Agency (DARPA) is an agency within the DOD that deals with military technologies. DARPA focuses on a 'new war' they

call a network war for cyber security. Understanding that nearly everything has a computer, including phone, television, watches, and military weapons systems, DARPA is utilizing a net-centric data strategy to develop mechanisms to thwart potential or actual cyber-attacks.

10.5 Summary

The big data tsunami created fertile ground for the conduct of research and related scholarship. It opened doors to healthcare research previously unimagined; it provided large data sets containing massive amounts of information with the potential of increasing knowledge and providing a proactive approach to healthcare. It also created a firestorm of change reflecting how access to data is accomplished. Big data is the automation of research. It also is an epistemological change that questions certain ethical morés; for example, just because we can access the data does not mean we should. Fordism changed the manufacturing world with a profound societal impact; big data is the new Fordism impacting society.

The societal and ethical impact of big data requires the attention of all disciplines. The impact, while requiring a priori decisions, will provide an unprecedented opportunity to influence healthcare and add to the global knowledge base and scholarly work. Big data has opened an abyss of opportunity to explore what is, hypothesize what could be, and provide methods to change practice, research and scholarship.

Big data is central to all areas of nursing research. Areas with the most prominent interface with other major healthcare initiatives, from an NIH perspective, include the precision medicine initiative seeking to further personalize a patient's health profile; the U.S. cancer moonshot, which has at its core the greater sharing and standardization of data supporting cancer research; and the Environmental influences of Child Health Outcomes (ECHO). We are truly entering the era of data-driven healthcare discovery and intervention.

References

- AACN. Position statement on nursing research. Washington, DC: American Association of Colleges of Nursing; 2006. <http://www.aacn.nche.edu/publications/position/nursing-research>
- Baca G. Legends of fordism: between myth, history, and foregone conclusions. *Soc Anal.* 2004;48(3):169–78. doi:10.3167/015597704782342393.
- Borgman CL, Bates MJ, Cloonan MV, Efthimiadis EN, et al. Social aspects of digital libraries. Final report to the National Science Foundation; 1996.
- Boyd D, Crawford K. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inform Commun Soc.* 2012;15(5):662–79. doi:10.1080/1369118X.2012.678878.
- Boyer EL. Scholarship reconsidered: priorities of the professoriate. Carnegie Foundation for the Advancement of Teaching; 1990.

- Brennan PF, Ponto K, Casper G, Tredinnick R, Broecker M. Virtualizing living and working spaces: proof of concept for a biomedical space-replication methodology. *J Biomed Inform.* 2015; doi:[10.1016/j.jbi.2015.07.007](https://doi.org/10.1016/j.jbi.2015.07.007).
- Deutsch T. Why is schema-on-read so useful? In: IBM big data and analytics hub. 2013; <http://www.ibmbigdatahub.com/blog/why-schema-read-so-useful>. Accessed 31 Jan 2016.
- Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genomic-wide expression patterns. *Proc Natl Acad Sci.* 1998;95:14863–8.
- Grady PA, Gough LL. Nursing science: claiming the future. *J Nurs Scholarsh.* 2015;47(6):512–21. doi:[10.1111/jnu.12170](https://doi.org/10.1111/jnu.12170).
- Hey T, Tanslery S, Tolle K. The fourth paradigm: data-intensive scientific discovery. Redmond, Washington: Microsoft Research; 2009.
- Hu CW, Kornblau SM, Slater, JH, Outub AA. Progeny clustering: a method to identify biological phenotypes nature; 2015. doi:[10.1038/srep12894](https://doi.org/10.1038/srep12894).
- Karr D. The flood of big data; 2012. <https://www.marketingtechblog.com/ibm-big-data-marketing>. Accessed 22 Dec 2015.
- Kitchin R (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences*. Sage: London
- Kling R. The internet and unrefereed scholarly publishing. *Annu Rev Inform Sci Technol.* 2004;38(1):591–631. doi:[10.1002/aris.1440380113](https://doi.org/10.1002/aris.1440380113).
- Kranzberg M. Technology and history: Kranzberg's laws. *Technol Cult.* 1986;27(3):544–60.
- Manyika J, Chui M, Brown B, Bughin, J, Dobbs R, Roxburgh C, Byers A. *Big data: the next frontier for innovation, competition, and productivity*. McKinsey & Company, The McKinsey Institute; 2011.
- McDonald L. Florence Nightingale and the early origins of evidence-based nursing. *Evid Based Nurs.* 2001;4(3):68–9. doi:[10.1136/ebn.4.3.68](https://doi.org/10.1136/ebn.4.3.68).
- National Institute of Nursing Research. 2015. <http://www.ninr.nih.gov/>. Accessed 17 Dec 2015.
- Pasqua J. Schema-on-read vs schema-on-write. In: MarkLogic; 2014. <http://www.marklogic.com/blog/schema-on-read-vs-schema-on-write/>. Accessed 17 Dec 2015.
- Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making. *Big Data.* 2013;1(1):BD51–9. doi:[10.1089/big.2013.1508](https://doi.org/10.1089/big.2013.1508).
- Rantz MJ, Banerjee TS, Cattoor E, Scott SD, Skubic S, Popescu M. Automated fall detection with quality improvement rewind to reduce falls in hospital rooms. *J Gerontol Nurs.* 2014;40(1):13–7. doi:[10.3928/00989134-20131126-01](https://doi.org/10.3928/00989134-20131126-01).
- Selanders LC, Crane PC. The voice of florence nightingale on advocacy. *Online J Issues Nurs.* 2012;17(1):1. doi:[10.3912/OJIN.Vol17No01Man01](https://doi.org/10.3912/OJIN.Vol17No01Man01).
- Tukey JW. The future of data analysis. *Ann Math Stat.* 1962;33(1):1–67. doi:[10.1214/aoms/1177704604](https://doi.org/10.1214/aoms/1177704604).
- Violino B. The hottest jobs in IT: training tomorrow's data scientists. *Forbes/Transformational Tech*; 2014.
- Wang W, Krishnan E. Big data and clinicians: a review on the state of the science. *JMIR Med Inform.* 2014;2(1):e1. doi:[10.2196/medinform.2913](https://doi.org/10.2196/medinform.2913).
- Westra BL, Delaney CW, Konicek D, Keenan G. Nursing standards to support the electronic health record. *Nurs Outlook.* 2008;56(5):258–266. e251. doi:[10.1016/j.outlook.2008.06.005](https://doi.org/10.1016/j.outlook.2008.06.005).

Case Study 10.1: Complexity of Common Disease and Big Data

Sandra Daack-Hirsch and Lisa Shah

Abstract Human development, health, and disease processes are the culmination of complex interactions among DNA sequences, gene regulation—epigenetics, and the environment. To truly create individualized interventions to address prevention and treatment, complex data systems that are integrated are needed. As an exemplar, this case study will explore the complexity of information and the vast sources of big data and related analytics needed to better understand causes of Type 2 Diabetes (T2D), which in turn will drive individualized interventions to reduce risk and better treat individuals. Personalized healthcare (also called precision medicine) is becoming a reality. President Obama announced the Precision Medicine Initiative in February 2015. Conceptually, precision medicine has been defined as prevention and treatment strategies that take individual differences into account to generate knowledge applicable to the continuum of health and disease. To that end, this case study describes current initiatives to assemble and analyze the vast and complex phenotypic, genetic, epigenetic, and exposure information generated or that will be generated by researchers and clinicians on individuals, within a specific clinical example of T2D.

Keywords Type 2 diabetes • Genetics/genomics • Exposome • Epigenetics • Omics • Big data • Personalized health care

10.1.1 Type 2 Diabetes (T2D) as a Significant Health Problem

Diabetes is a significant public health problem and its prevalence is increasing. As of 2012 an astounding 29.1 (9.3%) million Americans of all ages and racial/ethnic groups were affected by either Type 1 or Type 2 diabetes although, roughly 90–95% are Type 2. An estimated 8.1 million of those who have diabetes are undiagnosed (National Diabetes Information Clearinghouse [NDIC] 2014; Valdez et al. 2007) which potentiates the likelihood that they will also develop secondary health complications related to untreated diabetes (Klein Woolthuis et al. 2007). Diabetes is the leading cause of kidney failure, nontraumatic lower-limb amputation, heart disease, stroke, and new cases of blindness in the United States. In 2010 it was the seventh leading cause of death. As of 2012 medical expenses for people with diabetes (any form) were more than twice those for people without diabetes, costing Americans an estimated \$245 billion in direct and indirect costs (NDIC 2014).

S. Daack-Hirsch, Ph.D., R.N. (✉) • L. Shah, M.S.N., R.N.
The University of Iowa, Iowa city, IA, USA
e-mail: sandra-daack-hirsch@uiowa.edu

Type 2 diabetes (T2D) is a complex metabolic disease characterized by persistently elevated blood glucose caused by insulin resistance coupled with insulin deficiency. The substantial genetic component is thought to interact with environmental risk factors (plentiful diets and limited physical activity) to produce T2D. Several genetic variants have been associated with an increased risk to develop T2D (Wolfs et al. 2009), but the clinical validity of genetic variants alone to estimate diabetes risk remains limited, and genetic variants explain only a small portion of total risk variation (Burgio et al. 2015). While the staggering increase in prevalence of T2D is well documented, until recently the sharp increase in prevalence was largely thought to be driven by environmental factors experienced during adulthood (mainly an imbalance between energy intake and energy expenditure). However, there is also a sharp increase in T2D and obesity among persons under the age of twenty that challenges our understanding of risk factors (NDIC 2014). Just how genetic and environmental factors work in concert to produce the T2D phenotype remains unclear. Moreover, the environmental contribution to T2D is far more complex than an imbalance between energy intake and energy expenditure.

10.1.2 Factors Contributing to T2D

10.1.2.1 Genetics/Genomics

The genetic contribution to T2D has been established through research involving family history, twins studies, and genetic analysis. The lifetime risk for developing T2D in the Western world is reported to be between 7 and 10% (Burgio et al. 2015; Wolfs et al. 2009). Narayan et al. (2003) estimated that for individuals born in the United States in 2000, the lifetime risk for T2D is 1 in 3 for males and 2 in 5 for females. Family studies reveal that unaffected first-degree relatives of individuals with T2D have a two- to almost six-fold increase in risk to develop T2D over the course of a lifetime compared to people without a family history of T2D (Harrison et al. 2003; Valdez et al. 2007). The concordance rate for T2D among identical twins is high and is consistently reported to be greater than 50% in many populations (Medici et al. 1999; Newman et al. 1987; Poulsen et al. 2009). The fact that there are monogenetic (single gene) forms of diabetes (e.g., Maturity Onset Diabetes of the Young [MODY] and Permanent Neonatal Diabetes Mellitus [PNDM]) provides further evidence for a genetic role in the diabetes phenotype. Nevertheless, the majority of cases of diabetes are not one of the monogenetic forms; rather, T2D is a genetically complex disorder in which any number of genetic variants predispose an individual to develop the disease. Advances in genotyping technology have led to large-scale, population-based genetic studies to identify genetic variants (single nucleotide polymorphisms [SNPs]) associated with T2D. For example, the most recent studies have verified up to 65 SNPs to be associated with T2D (Morris et al. 2012; Talmud et al. 2015).

Discounting the rare monogenetic forms of diabetes and given the genetic and phenotypic heterogeneity of T2D¹ drawing firm conclusions about genotype—phenotype correlation using standard statistical analyses is difficult. To address statistical limitations, several efforts are underway to develop new algorithms to generate genetic risk scores (GRS) that predict T2D (Keating 2015; Talmud et al. 2015). As knowledge of the underlying genetic contribution grows, the prediction models improve. Mounting evidence shows that combining GRS and clinical risk factors (e.g., BMI, age, and sex) further improves the ability to detect incident cases (Keating 2015). However, the clinical utility of GRS remains problematic (Lyssenko and Laakso 2013). Developing risk prediction algorithms that combine GRS and phenotypic data for T2D is challenging in part due to the high heterogeneity in both the genetic factors and phenotypic elements of the disease. Moreover, genes and genetic variants at different locations in the genome (polygenic loci) that are associated with T2D are involved in multiple physiologic processes such as gluconeogenesis, glucose transport, and insulin homeostasis and many are also implicated in obesity (Burgio et al. 2015; Keating 2015; Slomko et al. 2012; Wolfs et al. 2009). To date, GRS use common genetic variants that have the strongest main effects. Other sources of genetic variance include rarer higher-penetrant variants, epigenetics, gene-gene and gene-environment interactions, and sex-specific genetic signals (Keating 2015; Lyssenko and Laakso 2013). The phenotype is also highly complex with patients presenting in various combinations of body type, age, family history, gestational diabetes, drug treatments, and comorbidities including obesity and metabolic syndrome. Detecting genetic differences is difficult when they are rare. Combining complex phenotype, interaction (gene x gene and or gene x environment), and gene variant information requires new data science approaches in order to leverage the complexity and create information that is clinically useful.

10.1.2.2 The Environment

In 2005 Wild coined the term “exposome” to describe the complementary environmental component of the gene-environment interaction indicative of complex traits and diseases (Wild 2005). As with the genetic component of T2D, the environmental component is also complex and plays a major role in the diabetes phenotype. Most of our knowledge of the T2D exposome is limited to the behavioral or modern living environment (Slomko et al. 2012). The modern living environment is characterized by increased access to low-cost, calorie-dense foods and increased sedentary lifestyle. The modern living environment is most amenable to intervention, and in fact interventions targeting diet and exercise are known to be effective in preventing or delaying the onset of T2D (Diabetes Prevention Program [DPP] Research Group 2002; Lindström et al. 2003; Venditti 2007). In the context of the modern living environment there is an emerging awareness of “unavoidable exposures” and their

¹Over 200 variants associated with Type 2 diabetes are recorded in the database of Genotypes and Phenotypes ([dbGaP] found at <http://www.ncbi.nlm.nih.gov/gap>).

connection to T2D (Slomko et al. 2012). These are exposures to man-made chemicals through ambient particles, water, food, and use of consumer or personal care products—some are found in plastics and resins. These chemicals are ubiquitous in the everyday environment at levels below standards set by the Environmental Protection Agency and other regulatory agencies. While a single exposure is not likely to cause harm, little is known about chronic low-level exposure and risk for disease. Burgio et al. (2015) summarized the growing evidence that suggests endocrine-disrupting chemicals such as brominated flame retardants and organochlorine pesticides, heavy metals, and pharmaceuticals (e.g., corticosteroids, antipsychotics, beta-blockers, statins, thiazide diuretics) may interfere with β -cell function and induce insulin resistance (Burgio et al. 2015, p.809; Diabetes.co.uk 2015).

There is also emerging evidence that the gut microbiota composition could affect risk for T2D. Gut microbiota are important for intestinal permeability, host metabolism, host energy homeostasis, and human toxicodynamics (how chemicals affect the body). Changes in microbiota composition that interfere with these functions can lead to increased activation of inflammatory pathways which in turn interferes with insulin signaling, increase in energy harvesting and fat storage in adipose tissue, and potentiate the effect of chemical exposure—all potential pathways to increase risk for metabolic syndrome, obesity and or T2D. For a more in depth review of gut microbiota and T2D refer to Burgio et al. (2015) and Slomko et al. (2012).

10.1.3 Epigenetics

10.1.3.1 Overview of Epigenetics

Epigenetics may explain how genetic and environmental factors work in concert to produce T2D. “Epigenetics is the study of heritable changes (either mitotically or meiotically) that alter gene expression and phenotypes, but are independent from the underlying DNA sequence ...”(Loi et al. 2013, p. 143). The epigenome is a series of chemical modifications (often referred to as tags or marks) that are superimposed on to the genome. In humans epigenetic modifications can either affect the proteins that are involved in the packaging of DNA into chromatin (known as chromatin modification), or directly attach to the DNA (e.g., DNA methylation). Epigenetic modification regulates gene expression by either activating (turning on) or deactivating (turning off) genes or segments of the DNA at given times (Genetics Learning Center 2014). Chromatin modification and DNA methylation are functionally linked to transcription and likely provide the mechanisms by which cells are programmed from one generation to the next. In other words, the epigenome activates genome in what is manifested as the phenotype.

During the pre-genomic era it was thought that disease and specific human traits were the direct result of variants in the DNA sequence (e.g., direct mutation of a single gene). However, very few diseases/traits are associated with only gene

variants. To varying degrees, other factors such as poverty, nutrition, stress, and environmental toxin exposures can also contribute to health or lack thereof; yet none fully explain susceptibility to disease or variations in human traits. Environmental and social signals such as diet and stress can trigger changes in gene expression without changing the sequence of the DNA (Heijmans et al. 2008; McGowan et al. 2009; Mathers et al. 2010; Radtke et al. 2011; Weaver et al. 2004). Some of these epigenetic tags are cell specific and differentiate phenotype at a cellular level with respect to cell type and function. In a differentiated cell, only 10 to 20% of the genes are active (Genetic Science Learning Center 2014). Some epigenetic tags are acquired and lost over the life course of an individual, and some tags are passed on from generation to generation and may take several generations to change.

10.1.3.2 Examples of Epigenetic Modification and T2D

While it is widely known and accepted that maternal nutrition is of paramount importance to the health and development of the offspring, the precise biologic mechanisms linking maternal nutrition to offsprings' wellbeing are just beginning to be understood. Epigenetic mechanisms may provide one such link. Evidence for epigenetic modification in the form of fetal programming can be found among individuals who were prenatally exposed to famine during the Dutch Hunger Winter in 1944–45. These individuals had less DNA methylation (hypomethylation) of the insulin-like growth factor 2 (IGF2) gene compared to their unexposed, same-sex siblings. IGF2 is a key factor in human growth and development. This epigenetic modification acquired in utero persisted throughout the children's lifetime (Heijmans et al. 2008) and has been associated with higher rates of T2D, obesity, altered lipid profiles, and cardiovascular disease (Schulz 2010) among these children (Burgio et al. 2015).

A number of recent studies report changes in methylation patterns of specific genes associated with T2D (Rönn et al. 2013; Zhang et al. 2013; Ling et al. 2008; Yang et al. 2011; Kulkarni et al. 2012; Yang et al. 2012; Hall et al. 2013; Ribel-Madsen et al. 2012). Studies also reveal differential methylation patterns in genes associated with T2D among those affected by T2D (Zhang et al. 2013; Ling et al. 2008; Yang et al. 2011, 2012; Kulkarni et al. 2012; Hall et al. 2013) and in tissue specific samples (pancreases and mitochondria). These types of studies provide evidence that genotypes (DNA sequences) and their regulation (epigenetic modifications) are important factors contributing to T2D and that the epigenome is modifiable providing targets for interventions.

The environmental exposures described above (“unavoidable exposures”) could also lead to changes in gut microbiota composition. In fact, changes in gut microbiota composition have been shown to interfere with epigenetic regulation of *FFAR3* gene in patients with T2D (Remely et al. 2014). *FFAR3* is normally expressed in the pancreatic β -cells and mediates an inhibition of insulin secretion by coupling with other proteins (National Center for Biotechnology Information [NCBI] 2015).

Interfering with the epigenetic regulation of FFAR3 would in turn lead to an inability to regulate insulin secretion appropriately.

10.1.3.3 Summary of Factors Contributing to T2D

T2D is the combination of biological contributing factors (genetics), environmental contributing factors (exposome), and the synthesis of biology and environment (epigenetics). Evolving epigenetic evidence suggests that epigenetic modifications could be important biomarkers for predicting risk, monitoring effectiveness of interventions, and targeting for therapy development to both prevent and treat T2D. Epigenetic patterns may serve as biomarkers connecting the exposome and genome (Fig. 10.1.1), thereby providing more comprehensive risk information for T2D. Unfavorable epigenetic modification may be reversed by lifestyle interventions, such as by modifying diet, increasing physical activity, and enriching the in utero environment. The rapid advances in genetic, exposome, and epigenetic sciences offer exciting possibilities for future discovery that will deepen our understanding of the complex balance between the environment and the genome, and how that balance influences health. Clearly an in-depth understanding T2D is largely dependent on big data and related advanced analytics.

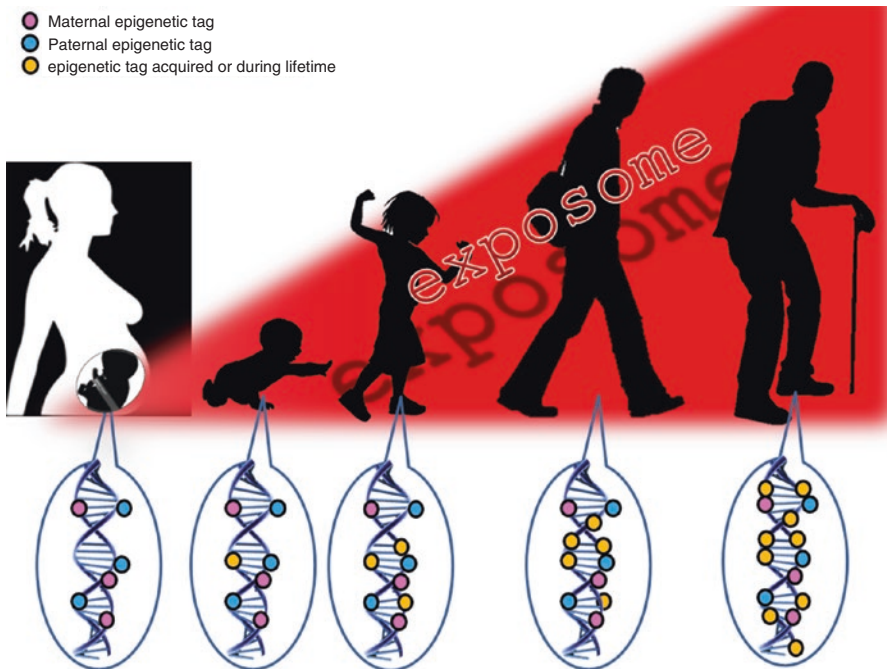


Fig. 10.1.1 Epigenetic modifications: the interaction of genetic and environmental risk factors over the life course

10.1.4 Current Initiatives to Leverage the Power of Big Data for Common Disease

10.1.4.1 Omics

Omics is the application of powerful high through-put molecular techniques to generate a comprehensive understanding of DNA, RNA, proteins, intermediary metabolites, micronutrients, and microbiota involved in biological pathways resulting in phenotypes. Scientists and informaticians are working on ways to integrate the layers of omic sciences and the exposome to better quantify an individual's susceptibility to diseases such as T2D and to capitalize on his or her inherent protections against disease (Slomko et al. 2012). These techniques would allow for massive amounts of genomic, epigenomic, exposure, and phenotypic data to be analyzed in concert in order to build more powerful prediction models and provide targets for the development of prevention and treatment modalities.

10.1.4.2 Clinical Genomic Resources

Several initiatives are underway to assemble the vast and complex phenotypic, genetic, epigenetic and exposure information pertaining to wellness and disease states. These initiatives will leverage health information that is currently generated or will be generated by researchers and clinicians on individuals.

ClinGen. (<http://clinicalgenome.org/>) is a project to develop standard approaches for sharing genomic and phenotypic data provided by clinicians, researchers, and patients through centralized databases, such as ClinVar—a National Database of Clinically Relevant Genetic Variants (CRGV). ClinGen investigators are working to standardize the clinical annotation and interpretation of genomic variants. Goals of ClinGen include:

- **Share** genomic and phenotypic data through centralized databases for clinical and research use
- **Standardize** clinical annotation and interpretation of variants
- Improve understanding of variation in **diverse populations**
- Develop **machine-learning algorithms** to improve the throughput of variant interpretation
- Implement **evidence-based expert consensus** for curation of clinical validity
- Assess the '**medical actionability**' of genes and variants to support their use in clinical care systems
- **Disseminate** the collective knowledge/resources and ensure EHR interoperability (<http://www.genome.gov/27558993>)

Currently ClinGen efforts are focused on cardiovascular disease, pharmacogenomics, hereditary (germline) cancer, somatic cancer, and inborn errors of metabolism. However, knowledge generated on structure and process will serve as a template for approaching other diseases.

eMERGE. The Electronic Medical Records and Genomics (eMERGE) Network is a National Institutes of Health (NIH)-organized and funded consortium of U.S. medical research institutions. The eMERGE Network brings together researchers from leading medical research institutions across the country to conduct research in genomics, including discovery, clinical implementation and public resources. eMERGE was announced in September 2007 and began its third and final phase in September 2015 (<http://www.genome.gov/27558993>). The Network is comprised of six workgroups (see Table 10.1.1).

The primary goal of the eMERGE Network is to develop, disseminate, and apply approaches to research that combine biorepositories with electronic medical record (EMR) systems for genomic discovery and genomic medicine implementation research. In addition, the consortium includes a focus on social and ethical issues such as privacy, confidentiality, and interactions with the broader community (eMERGE <https://emerge.mc.vanderbilt.edu/>).

PhenX. One of the limitations in being able to interpret findings from genome-wide association (GWA) studies is lack of uniform phenotypic descriptions and measures. For example, hundreds of associations between genetic variants and diabetes have been identified. However, most GWA studies have had relatively few phenotypic and exposure measures in common. Development and adoption of standard phenotypic and exposure measures could facilitate the creation of larger and more comprehensive datasets with a variety of phenotype and exposure data for cross-study analysis, thus increasing statistical power and the ability to detect associations of modest effect sizes and gene-gene and gene-environment interactions (<http://www.genome.gov/27558993>). PhenX was developed in recognition of the need for standard phenotypic and exposure measures, particu-

Table 10.1.1 eMerge workgroup summary. eMERGE <https://emerge.mc.vanderbilt.edu/>

Workgroup	Primary function of the workgroup
Phenotyping workgroup	Creation, validation, and execution of phenotype algorithms across the network and beyond
Genomics workgroup	Create a single and uniform data set for all individuals genotyped across the network
Return of Results (RoR)/ Ethical, Legal, and Social Implications (ELSI) workgroup	Define an initial set of variants that are potentially useful in clinical practice for purposes such as assessment of genetic risk for complex disorders or selection or dosing of drugs, focusing on common disease risk variants and pharmacogenetic variants for which we expect to have data; assess the levels of evidence supporting these variants and consider the cost and benefit of incorporating them into patient care
EHR Integration (EHRI) workgroup	Develop standards and methods for incorporating genomic data into the EHR and optimizing the utilization of that data for patients and physicians
Clinical annotation workgroup	Work with Central Sequencing and Genotyping Centers to make pathogenic/non-pathogenic calls on samples
Outcomes workgroup	Study economic and non-economic outcomes of genomic research integrated into healthcare

larly as related to GWA studies. The National Human Genome Research Institute (NHGRI) initiated the [PhenX Toolkit](#) in 2006 with the goal of identifying and cataloguing 15 high-quality, well-established, and broadly applicable measures for each of 21 research domains (diabetes is one of these) for use in GWA studies and other large-scale genomic research (www.phenxtoolkit.org).

Roadmap Epigenetics Mapping Consortium. The National Institute of Health (NIH) Roadmap Epigenetics Mapping Consortium was created in an effort to understand epigenetic modifications and how these interact with underlying DNA sequences to contribute to health and disease. The project will provide publically available epigenetic maps on normal human tissues, support technology development, and provide funding in epigenetics research (National Institutes of Health 2015; Slomko et al. 2012).

Precision Medicine/Personalized Healthcare. Precision Medicine/personalized healthcare is a medical model that proposes to customize healthcare by incorporating medical decisions, practices, and products that are based on individual variability in genes, environment, and lifestyle. The potential ability of applying this concept more broadly has been dramatically improved by the recent development of large-scale biologic databases described above. The Precision Medicine Initiative Cohort Program proposes to:

- Identify genomic variants that affect drug response
- Assess clinical validity of genomic variants associated with disease
- Identify biomarkers that are early indicators of disease
- Understand chronic diseases and best management strategies
- Understand genes/pathways/factors that protect from disease
- Assess how well novel cellphone-based monitors of health work
- Evaluate the ability of EHRs to integrate research data
- Learn and apply new ways of engaging participants in research
- Develop methodology for data mining and statistical analysis (<https://www.nih.gov/precision-medicine-initiative-cohort-program>)

10.1.5 Scope and Practice of Genetics/Genomics Nursing

The American Nurses Association in collaboration with the International Society of Nurses in Genetics provides an excellent resource for nurses interested in clinical genetics and nursing, the *Genetics/Genomics Nursing: Scope and Standards of Practice*, 2nd Edition (2016). This resource summarizes the role of nurses in genetics/genomics, which focuses on the actual and potential impact of genetic/genomic influences on health. Genetics/genomics nurses educate clients and families on genetic/genomic influences that might impact their health and intervene with the goals of optimizing health, reducing health risks, treating disease, and promoting wellness. This practice depends upon research and evidence-based practice, inter-professional collegiality and collaboration with genetics/genomics professionals and other healthcare professionals to provide quality patient care.

10.1.6 Conclusion

In conclusion, T2D is an increasingly common and complex disorder with genome, exposome, and epigenome factors contributing to the widely variable phenotype. Initiatives in precision medicine propose to customize healthcare by integrating data and information pertaining to individual variability in genes, environment, and lifestyle and interpreting this information to inform medical decisions, practices, and products that prevent, delay, and more effectively treat individuals who are at risk or have T2D. While many of our current initiatives build the evidence base needed to guide clinical practice for the individual, society also needs to be mindful of the social inequalities of opportunity including education, environmental quality, and access, not only to health care but to nutritious food, recreation, and community supports that contrive health and disease. These social determinants are part of the individual's exposome, and yet are often beyond the control of the individual. Finally, motivating individuals at higher risk to engage in lifestyle changes to reduce their risk for T2D remains challenging. Communicating risk information about T2D is further complicated by how a person personalizes and rationalizes his or her risk to develop it (Shah et al. in press; Walter and Emory 2005). Knowing about genetic risk is not enough to motivate people to change behaviors (Grant et al. 2013). An important knowledge gap to fill is our understanding of how people at increased risk for T2D come to understand and manage behaviors to reduce their risk for disease. Understanding a person's beliefs may facilitate effective collaboration with healthcare providers, and improve risk reduction education using a truly comprehensive personalized approach.

References

- American Nurses Association. Genetics/genomics nursing: scope and standards of practice. 2nd ed. Washington, DC: ANA; 2016.
- Bell G (2009). Forward in *The Fourth Paradigm: Data-Intensive Scientific Discovery* (p.xv). Redmond, WA, Microsoft.
- Burgio E, Lopomo A, Migliore L. Obesity and diabetes: from genetics to epigenetics. *Mol Biol Rep.* 2015;42:799–818. doi:[10.1007/s11033-014-3751-z](https://doi.org/10.1007/s11033-014-3751-z).
- Diabetes.co.uk the Global Diabetes Community. Drug induced diabetes. 2015. <http://www.diabetes.co.uk/drug-induced-diabetes.html>. Accessed 30 Jan 2016.
- Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Nathan DM. Diabetes prevention program research group 2002. Reduction in the incidence of Type 2 diabetes with lifestyle intervention or Metformin. *N Engl J Med.* 2002;346(6):393–403. doi:[10.1056/NEJMoa012512](https://doi.org/10.1056/NEJMoa012512).
- Diabetes Prevention Program [DPP] Research Group 2002. The diabetes prevention program (DPP): description of lifestyle intervention. *Diabetes Care.* 2002;25(12):2165–71. doi:[10.2337/diacare.25.12.2165](https://doi.org/10.2337/diacare.25.12.2165).
- Genetic Science Learning Center. Epigenetics. 2014. <http://learn.genetics.utah.edu/content/epigenetics/>. Accessed 30 Jan 2016.
- Grant RW, O'Brien KE, Waxler JL, Vassy JL, Delahanty LM, Bissett LG, Green RC, Stember KG, Guiducci C, Park ER, Florez JC, Meigs JB. Personalized genetic risk counseling to motivate diabetes prevention: a randomized trial. *Diabetes Care.* 2013;36(1):13–9. doi:[10.2337/dc12-0884](https://doi.org/10.2337/dc12-0884).

- Hall E, Dayeh T, Kirkpatrick CL, Wollheim CB, Dekker Nitert M, Ling C. DNA methylation of the glucagon-like peptide 1 receptor (GLP1R) in human pancreatic islet. *BMC Med Genet.* 2013;14:76. doi:10.1186/1471-2350-14-76.
- Harrison TA, Hindorff LA, Kim H, Wines RC, Bowen DJ, McGrath BB, Edwards KL. Family history of diabetes as a potential public health tool. *Am J Prev Med.* 2003;24(2):152–9. doi:10.1016/S0749-3797(02)00588-3.
- Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, Slagboom PE, Lumey LH. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings from the National Academy of Science.* 2008;105(44):17046–9. doi:10.1073/pnas.0806560105.
- Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet.* 2003;33(Suppl):245–54. doi:10.1038/ng1089.
- Keating BJ. Advances in risk prediction of type 2 diabetes: integrating genetic scores with Framingham risk models. *Diabetes.* 2015;64(5):1495–7. doi:10.2337/db15-0033.
- Klein Woolthuis EP, de Grauw WJ, van Gerwen WH, van den Hoogen HJ, van de Lisdonk EH, Metsemakers JF, van Weel C. Identifying people at risk for undiagnosed Type 2 diabetes using the GP's electronic medical record. *Fam Pract.* 2007;24(3):230–6. doi:10.1093/fampra/cmm018.
- Kulis M, Esteller M. DNA methylation and cancer. *Adv Genet.* 2010;70:27–56. doi:10.1016/B978-0-12-380866-0.60002-2.
- Kulkarni SS, Salehzadeh F, Fritz T, Zierath JR, Krook A, Osler ME. Mitochondrial regulators of fatty acid metabolism reflect metabolic dysfunction in Type 2 diabetes mellitus. *Metabolism.* 2012;6(2):175–85. doi:10.1016/j.metabol.2011.06.014.
- Lindström J, Louheranta A, Mannelin M, Rastas M, Salminen V, Eriksson J, Uusitupa M, Tuomilehto J. Finnish Diabetes Prevention Study group 2003. The Finnish Diabetes Prevention Study (DPS): lifestyle intervention and 3-year results on diet and physical activity. *Diabetes Care.* 2003;26(12):3230–6. doi:10.2337/diacare.26.12.3230.
- Ling C, Del Guerra S, Lupi R, Rönn T, Granhall C, Luthman H, Masiello P, Marchetti P, Groop L, Del Prato S. Epigenetic regulation of PPARGC1A in human Type 2 diabetic islets and effect on insulin secretion. *Diabetologia.* 2008;51(2):615–22. doi:10.1007/s00125-007-0916-5.
- Lysenko V, Laakso M. Genetics screening and the risk for Type 2 diabetes: worthless or valuable? *Diabetes Care.* 2013;36(Suppl(2)):S120–6. doi:10.2337/dcS13-2009.
- Loi M, Del Savio L, Stupka E. Social epigenetics and equality of opportunity. *Public Health Ethics.* 2013;6(2):142–53. doi:10.1093/phe/pht019.
- McGowan PO, Sasaki A, D'Alessio AC, Dymov S, Labonté B, Szyf M, Turecki G, Meaney MJ. Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. *Nat Neurosci.* 2009;12(3):342–8. doi:10.1038/nn.2270.
- Mathers JC, Strathdee G, Reltón CL. Induction of epigenetic alterations by dietary and other environmental factors. *Adv Genet.* 2010;71:3–39. doi:10.1016/B978-0-12-380864-6.00001-8.
- Medici F, Hawa M, Ianari A, Pyke DA, Leslie RD. Concordance rate for type II diabetes mellitus in monozygotic twins: actuarial analysis. *Diabetologia.* 1999;42(2):146–50. doi:10.1007/s001250051132.
- Morris AP, Voight BF, Teslovich TM, et al. Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet.* 2012;44:981–90. doi:10.1038/ng.2383.
- National Diabetes Information Clearinghouse [NDIC]. National diabetes statistics. 2014. <http://diabetes.niddk.nih.gov/dm/pubs/statistics/#Prevention>. Accessed 27 Jan 2016.
- National Center for Biotechnology Information (NCBI). FFAR3 free fatty acid receptor 3 [*Homo sapiens* (human)]. 2015. <http://www.ncbi.nlm.nih.gov/gene/2865>. Accessed 30 Jan 2016.
- National Institutes of Health. The National Institutes of Health (NIH) roadmap epigenomics mapping consortium. 2015. <http://www.roadmapepigenomics.org>. Accessed 30 Jan 2016.

- Narayan KM, Boyle JP, Thompson TJ, Sorenson SW, Williamson DF. Lifetime risk for diabetes mellitus in the United States. *JAMA*. 2003;290(14):1884–90. doi:10.1001/jama.290.14.1884.
- Newman B, Selby JV, King MC, Slemenda C, Fabsitz R, Friedman GD. Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins. *Diabetologia*. 1987;30(10):763–8. doi:10.1007/BF00275741.
- Poulsen P, Grunnet LG, Pilgaard K, Storgaard H, Alibegovic A, Sonne MP, Carstensen B, Beck-Nielsen H, Vaag A. Increased risk of Type 2 diabetes in elderly twins. *Diabetes*. 2009;58(6):1350–5. doi:10.2337/db08-1714.
- Puumala SE, Hoyme HE. Epigenetics in pediatrics. *Pediatr Rev*. 2015;36:14–21. doi:10.1542/pir.36-1-14.
- Radtke KM, Ruf M, Gunter HM, Dohrmann K, Schauer M, Meyer A, Elbert T. Transgenerational impact of intimate partner violence on methylation in the promoter of the glucocorticoid receptor. *Transl Psychiatry*. 2011;1:21. doi:10.1038/tp.2011.21.
- Remely M, Aumueller E, Merold C, Dworzak S, Hippe B, Zanner J, Pointner A, Brath H, Haslberger AG. Effects of short chain fatty acid producing bacteria on epigenetic regulation of FFAR3 in type 2 diabetes and obesity. *Gene*. 2014;537:85–92. doi:10.1016/j.gene.2013.11.081.
- Ribel-Madsen R, Fraga MF, Jacobsen S, Bork-Jensen J, Lara E, Calvanese V, Fernandez AF, Friedrichsen M, Vind BF, Højlund K, Beck-Nielsen H, Esteller M, Vaag A, Poulsen P. Genome-wide analysis of DNA methylation differences in muscle and fat from monozygotic twins discordant for type 2 diabetes. *PLoS One*. 2012;7(12):e51302. doi:10.1371/journal.pone.0051302.
- Rönn T, Volkov P, Davegårdh C, Dayeh T, Hall E, Olsson AH, Nilsson E, Tornberg A, Dekker Nitert M, Eriksson KF, Jones HA, Groop L, Ling C. A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue. *PLoS Genet*. 2013;9(6):e31003572. doi:10.1371/journal.pgen.1003572.
- Schulz LC. The Dutch hunger winter and the developmental origins of health and disease. *Proceedings from the National Academy of Science*. 2010;107(39):16757–8. doi:10.1073/pnas.1012911107.
- Shah L, Perkhounkova Y, Daack-Hirsch S. Evaluation of the perception of risk factors for Type 2 diabetes instrument (PRF-T2DM) in an at-risk, non-diabetic population. *J Nurs Meas*. In press.
- Slomko H, Heo HJ, Einstein FH. Minireview: epigenetics of obesity and diabetes in humans. *Endocrinology*. 2012;153(3):1025–30. doi:10.210/en.2011-1759.
- Strahl BD, Allis CD. The language of covalent histone modifications. *Nature*. 2000;403(6765):41–5. doi:10.1038/47412.
- Talmud PJ, Cooper JA, Morris RW, Dudbridge F, Shah T, Engmann J, Dale C, White J, McLachlan S, Zabaneh D, Wong A, Ong KK, Gaunt T, Holmes MV, Lawlor DA, Richards M, Hardy R, Kuh D, Wareham N, Langenberg C, Ben-Shlomo Y, Wannamethee SG, Strachan MW, Kumari M, Whittaker JC, Drenos F, Kivimaki M, Hingorani AD, Price JF, Humphries SE. Sixty-five common genetic variants and prediction of Type 2 diabetes. *Diabetes*. 2015;64(5):1830–40. doi:10.2337/db14-1504.
- Valdez R, Yoon PW, Liu T, Khoury MJ. Family history and prevalence of diabetes in the U.S. population: the 6-year results from the National Health and Nutrition Examination Survey (1999–2004). *Diabetes Care*. 2007;30(10):2517–22. doi:10.2337/dc07-0720.
- Venditti EM. Efficacy of lifestyle behavior change programs in diabetes. *Curr Diab Rep*. 2007;7(2):123–7. doi:10.1007/s11892-007-0021-7.
- Walter FM, Emery J. Coming down the line—patients’ understanding of their family history of common chronic disease. *Ann Fam Med*. 2005;3(5):405–14. doi:10.1370/afm.368.
- Weaver IC, Cervoni N, Champagne FA, D’Alessio AC, Sharma S, Seckl JR, Dymov S, Szyf M, Meaney MJ. Epigenetic programming by maternal behavior. *Nat Neurosci*. 2004;8:847–54. doi:10.1038/nn1276.
- Wild CP. Complementing the genome with an ‘exposome’: the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomark Prev*. 2005;14(8):1847–50. doi:10.1093/mutage/gen061.

- Wolfs MG, Hofker MH, Wijmenga C, van Haefte TW. Type 2 diabetes mellitus: new genetic insights will lead to new therapeutics. *Curr Genomics*. 2009;10(2):110–8. doi:[10.2174/138920209787847023](https://doi.org/10.2174/138920209787847023).
- Yang BT, Dayeh TA, Kirkpatrick CL, Taneera J, Kumar R, Groop L, Wollheim CB, Nitert MD, Ling C. Insulin promoter DNA methylation correlates negatively with insulin gene expression and positively with HbA(1c) levels in human pancreatic islets. *Diabetologia*. 2011;54(2):360–7. doi:[10.1007/s00125-010-1967-6](https://doi.org/10.1007/s00125-010-1967-6).
- Yang BT, Dayeh TA, Volkov PA, Kirkpatrick CL, Malmgren S, Jing X, Renström E, Wollheim CB, Nitert MD, Ling C. Increased DNA methylation and decreased expression of PDX-1 in pancreatic islets from patients with type 2 diabetes. *Mol Endocrinol*. 2012;26(7):1203–12. doi:[10.1210/me.2012-1004](https://doi.org/10.1210/me.2012-1004).
- Zhang Y, Kent 2nd JW, Lee A, Cerjak D, Ali O, Diasio R, Olivier M, Blangero J, Carless MA, Kissebah AH. Fatty acid binding protein 3 (fabp3) is associated with insulin, lipids and cardiovascular phenotypes of the metabolic syndrome through epigenetic modifications in a Northern European family population. *BMC Med Genet*. 2013;6:9. doi:[10.1186/1755-8794-6-9](https://doi.org/10.1186/1755-8794-6-9).