Chapter 3
Data and Computation: A Contemporary Landscape

Ida Sim and Marina Sirota

After reading this chapter, you should know the answers to these questions:

- What type of data can be leveraged for medical research and care?
- How do we know and learn about the world through data and computation?
- What computational infrastructures currently exist to support research discovery and clinical care?
- What are artificial intelligence and machine learning and how are they related?
- What types of knowledge representation exist?
- What are open challenges in the field moving forward?

Understanding the World Through Data and Computation

Data has been called the "new oil" [1] or likened to "sunlight" [2] in its ubiquity and importance. Yet no one goes to medical school to learn data; one goes to medical school to learn what's needed to diagnose, treat, and care for people. What then is the role of data in biomedicine? Ackoff [3] is often credited with positing the data-information-knowledge continuum, in which data are raw observations, information is data in context, and knowledge is an understanding about the world that is useful for explaining, predicting, and guiding future action. Knowledge—what we learn in medical school—may be explicit and codifiable (e.g., guidelines, textbooks), tacit and not codifiable (e.g., expertise, heuristics), or process knowledge (e.g., how to

I. Sim (✉) · M. Sirota
University of California San Francisco, San Francisco, CA, USA
e-mail: ida.sim@ucsf.edu; marina.sirota@ucsf.edu

T. A. Cohen et al. (eds.), *Intelligent Systems in Medicine and Health*, Cognitive Informatics in Biomedicine and Healthcare,
https://doi.org/10.1007/978-3-031-09108-7_3

remove a gallbladder). Here is a clinical example. An observation that a patient's Hemoglobin A1c (**HbA1c)** is 8.2% is data; that this HbA1c of 8.2% is above the normal range is information, i.e., data in context; that a high HbA1c is associated with increased risk of adverse cardiovascular outcomes is knowledge. Knowledge is used, along with data and information about specific patients or populations, to guide actions in clinical care and population health respectively.

In recent years, machine learning and other computational approaches have powered a new path to transforming data into knowledge. But of course, biomedicine had been generating knowledge from data well before the modern era of computing. The dominant epistemology of clinical medicine—"the investigation of what distinguishes justified belief from opinion" [4]—became increasingly grounded in the scientific method starting at the turn of the twentieth century, progressed as a result of the 1910 Flexner Report [5] to formalized teaching of physiology and biochemistry in medical school (See Chap. 16), and culminated with the tenets of evidence-based medicine (EBM) as described by Guyatt and others in 1992 [6]. EBM is marked by scrupulous attention to experimental sources of bias that may cloud attempts to distinguish "justified belief from opinion." The randomized controlled trial (RCT), which controls for both known and unknown confounders through randomization, was held up as the gold standard for resolving questions of causation, sitting atop the evidence hierarchy save for the aggregation of RCTs in meta-analysis (Fig. 3.1).

However, this classical formulation of EBM addresses only questions of causation (does X cause Y). RCTs are not an appropriate study design for other types of questions central to clinical care [7], including description of **natural history** (what happens to people with Stage 5 lung cancer), **classification** (does this

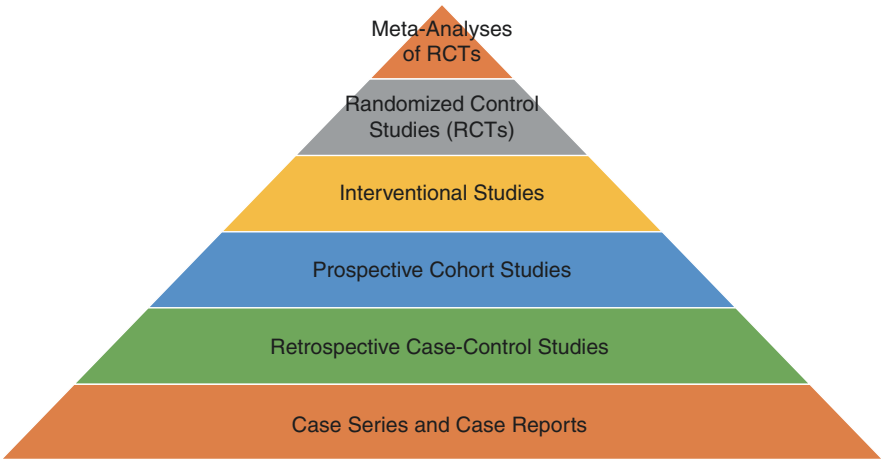

Fig. 3.1 Hierarchy of evidence according to evidence-based medicine

patient belong in (i.e., is classifiable into) the group of patients with Type 2 diabetes), **prediction** (how long will this patient with Stage 5 cancer live), and **explanation** (how does a high HbA1C result in elevated cardiovascular risk). An expanded version of EBM now addresses these other epistemological tasks using other study designs such as case control studies, prospective cohort studies, and prognostic rules [8].

Evidence : data + "study" design + analysis

Evidence is generated from data collected according to some study protocol (e.g., for an RCT, cohort study, or systematic review) and analyzed through biostatistical methods (e.g., intention-to-treat analysis for RCTs). The analyses generate findings which are used to support claims of knowledge (e.g., dexamethasone reduces 28-day mortality in some hospitalized patients with COVID-19 [9]). A particular claim of knowledge is justified beyond simple belief based on the evidentiary strength of the study design and analytic method. The claim that dexamethasone is efficacious for COVID-19 as supported by a well-conducted RCT can be contrasted with a belief in some circles of hydroxychloroquine's efficacy.

The contemporary landscape of biomedical epistemology is in tension and flux. While much of clinical research is still firmly embedded in traditional EBM approaches to generating evidence and knowledge, new computational approaches analyze vast amounts of data using "study designs" or algorithms that are wholly different from how clinical researchers and clinicians have been taught to know the world. Logistic regression and various machine learning algorithms are both analytic methods applied to data to generate evidence for claims of knowledge. These two ways of knowing [10]—EBM and data science—are complementary and can both be advanced with contemporary computational capabilities. This chapter reviews the foundations of data and computation as an underpinning to the following chapters.

Types of Data Relevant to Biomedicine

There are many broad classes of data relevant to biomedicine and healthcare, including Electronic Health Records (EHR), -omics, imaging, mobile and social media, environmental, public health, and clinical research data. The EHR captures patient information including demographics, diagnosis codes, lab test results, medications, allergies, and clinical notes generated from the provision of health care. While these data are originally collected for clinical and reimbursement purposes, they provide an incredible opportunity to mine and apply machine learning techniques for predicting disease risk or understanding disease better. These data have been used widely to predict patient outcomes such as hospital readmission rate [11] or

pregnancy outcomes [12]. Other clinical datasets include MIMIC-IV [13], a large, single-center database containing information relating to patients admitted to critical care units at a large tertiary care hospital. MIMIC is a rare example of a large clinical dataset available for use by the broader research community. There are efforts in clinical trials data sharing through repositories such as ImmPort [14] and Vivli [15]. Finally clinical imaging is another field with many opportunities to apply advanced machine learning and predictive modeling techniques for diagnostic purposes, as further described in Chap. 12.

Genomic and other molecular profiling technologies allow us to extract large amounts of data from patient samples, elucidating previously unknown factors involved in disease, such as drug targets or disease biomarkers. Much of the data from these types of experiments are publicly available. For instance, gene expression data are hosted in the Gene Expression Omnibus (GEO) [16] that as of July 2021, contains data on over 4.5 million samples and over 150,000 experiments. These data are very rich, capturing a number of different disease areas. With the technologies getting cheaper and more advanced, many of the transcriptomic studies now capture expression on a single cell level. dbGAP [17] and Short Read Archive (SRA) both house sequencing data with additional security for ensuring patient privacy. There are also disease-specific databases such as the Cancer Genome Atlas (TCGA) [18] that contains molecular measurements on more than 10,000 cancer samples and adjacent normal controls including transcriptomics, genetics, methylation and proteomics. The Preterm Birth Data Repository [19] is another example of a data repository, which as of July 2021 hosted over 45 molecular studies relating to pregnancy outcomes with a focus on preterm birth. A more in-depth description of applications of artificial intelligence to molecular measurements as part of the field of translational bioinformatics can be found in Chap. 14.

Clinical and molecular datasets can furthermore be enhanced by public health data such as The National Health and Nutrition Examination Survey (NHANES). NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States and uniquely combines interviews and physical examinations. CalEnviroscreen [20] is a database that captures environmental exposures across the state of California. Birth and death records (e.g., OSHP [21]) have been used extensively for research purposes. For instance in our own work, we have integrated the environmental exposure data from the CalEnviroscreen together with birth records information in order to identify arsenic and nitrate as water contaminants that are associated with preterm birth [22]. Finally in the last several years, mobile/social media data such as actigraphy, Twitter, and smartwatch data has been used to improve disease diagnosis (e.g., of atrial fibrillation [23]), monitor symptoms [24], and drive health behavior change [25, 26]. Newer modalities of data acquisition including Ecological Momentary Assessments (EMAs) [27]

that prompt users for their behaviors and experiences in real time in their natural environments are offering an unprecedented view into people's lived experience of health and disease.

The truth, however, is that there is no such thing as "health" and "non-health" data: all data can have implications for health. For example, individual-level data such as your online purchases, social media, geolocation, financial and criminal record data can be mined for predictors of health risk and health status. Environmental and population-level data such as block-level air pollution and noise [28], and voting patterns in your state [29], could be as predictive for health as traditional EHR data. The boundaries dividing health from general societal data and computing infrastructure are increasingly porous.

Knowing Through Computation

The explosive availability of Big Data—distinguished by high Velocity (speed of data generation), Volume, and Variety—enables new levels of data-driven reasoning, of which there are two major flavors. **Abductive reasoning** as originally coined by Pierce in 1955 [30] can be characterized as a cyclical process of *generating* possible explanations or a set of hypotheses that are able to account for the available data (see also the similar discussion of these concepts as they apply to human reasoning in Chap. 5). More recently, the term abductive reasoning has been expanded to the notion of "Inference to the Best Explanation" [31], by which a hypothesis or theory is arrived at that *best explains* the available data. Over time, clinical research using traditional statistics also endeavors to arrive at a "best explanation." A study postulates a hypothesis, data is collected and analyzed drawing on deep domain expertise, and the null hypothesis is accepted or rejected thus arriving at a provisional explanation of the observed data. Randomized controlled trials are a type of study design that controls for known and unknown confounders to strengthen a claim of causation, yielding a "best explanation" that can be contravened by other or subsequent trials. In computation, case-based reasoning is a classic example of abductive decision support systems, which are nowadays overshadowed by inductive machine learning approaches.

Inductive reasoning involves an inferential process from the observed data to account for the unobserved. It is a process of generating possible conclusions based on available data. The power of inductive reasoning lies in its ability to allow us to go beyond the limitations of our current evidence or knowledge to novel conclusions about the unknown. **Machine learning**—computer algorithms that find and apply patterns in (huge amounts of) data—is quintessential inductive reasoning. Subtypes include **classification**, **prediction**, **causal reasoning,** and **modeling** (Box 3.1).

Box 3.1 Examples of Inductive Reasoning

Classification: *inferring which class an instance belongs to based on classes of observed instances. E.g., a diagnostic decision support system "classifies" a given patient to a "disease" based on the similarity of their symptoms to the symptoms of prior patients known to have that disease.*

Prediction: *inferring a future state based on past data. E.g., a clinical decision support system predicts whether a patient will require hospitalization based on historical hospital admissions data.*

Causation: *inferring whether X caused Y. E.g., a deep neural network running over a clinical data warehouse is used to discover whether Drug X causes a Side Effect Y.*

Modeling*: simulating the components, relationships, and actions within a biomedical or health system to explain, explore, or predict E.g., a discrete event model of endocrine feedback for a disease.*

The combination of Big Data and machine learning is fueling a transformation in computational reasoning. Coupled with advances in cloud, social, mobile and other technologies, a new frontier is opening up for what computers can do with and for humans in health and biomedicine.

Motivational Example

Now that we have an overview of biomedical data and computation, we present and deconstruct an example clinical case (Box 3.2) to illustrate high-level issues and challenges that will shape the near future of data and computation.

Box 3.2 Illustrative Case

Andre is a 47-year-old man with mild Type 2 diabetes. He was returning from a business trip overseas when he felt short of breath, out of sorts, and had occasional sharp chest pains. He signed onto a telehealth service offered through his employer. The telehealth service's chatbot interviewed Andre, using an avatar that was Hispanic, as Andre is. After an initial set of questions, the chatbot handed over the case to a human physician, who conducted a video consultation with Andre while reviewing his electronic health record data along with his respiratory rate, body temperature, oxygen saturation and other data from his smartwatch. The physician recommended that Andre get evaluated in person at the nearest Emergency Room (ER). Andre is getting worried. On his way to the ER, Andre asks Siri what he might have. Siri tells him scary diagnoses like pneumonia, and something called pulmonary embolism. Siri explains that pulmonary embolism is when a blood clot forms in a

leg after prolonged sedentariness (like a long flight) and breaks off to the lungs causing chest pain and shortness of breath.

At the ER, Andre was first seen by a resident physician-in-training who ordered multiple tests including labwork, a chest x-ray, and a chest CT. Based on those data, a decision support system ran predictive models that resulted in a ranked list of differential diagnoses, with an intermediate probability for pulmonary embolism. The resident presented the case to Dr. Jackson, the attending ER physician. After reviewing the data and output, Dr. Jackson went to talk with Andre and examine him. She noticed crackles in the lungs, an S3, a prominent right-sided cardiac lift and elevated jugular venous pressure. On further questioning, Andre mentioned that he had had a "bad cold" about 1 month before and had been feeling unwell even before the business trip. Suspecting biventricular failure from viral myopericarditis, Dr. Jackson ordered an echocardiogram and admitted Andre.

Andre is fortunate to have convenient timely access to "virtual-first" care through his employer. 9% of Americans have no health insurance [32] at all while 43% are underinsured [33]. When health care moved onto virtual platforms during the SARS-CoV-2 pandemic, marginalized populations had reduced access to health care due to lack of technology and/or technology literacy [34], adding "digital determinants of health" to the causes of health inequities (Chaps. 13 and 18). As with general consumer technology, chatbot services are increasingly common in health. Chapter 9 reviews natural language processing (NLP) and other computational issues underlying dialog systems. Culturally concordant avatars, language, and user interactions are needed to establish belonging and trust with digital interactions for all peoples (Chap. 18). Central to this book on cognitive informatics is the importance of a smooth handoff between computational and human care: the decision to refer to Andre to the ER is one that should involve a human, who in this case was able to access and review Andre's EHR and wearable data to get a better view of his overall status. The ability to access such data in real time requires health data interoperability encompassing network computing, data standards, and sociotechnical data sharing mechanisms. Siri and the decision support system in the ER illustrate the exciting possibilities of automated reasoning. Early diagnostic systems dating from the 1970s include INTERNIST-1 and **MYCIN** (Chap. 2). Simpler systems, such as the Modified Early Warning System (MEWS) for scoring physiologic observations to predict sepsis [35], have been widely used in clinical practice, and have evolved to machine-learning based models with better performance (Chap. 10). Advances in image recognition have given rise to imaging decision support systems such as for detecting pulmonary embolism (Chap. 12).

Andre's case illustrates the importance of framing clinical decision support not as a solely computational task but as one of human/AI collaboration requiring a human-in-the-loop approach. The ER resident who first evaluated Andre likely had

premature closure [36] on the potential diagnosis of pulmonary embolism (PE) and collected data (e.g., chest CT) with PE in mind while not pursuing other potential diagnoses. When presented with this restricted set of data, the decision support system backs up the resident's diagnostic hunch. Dr. Jackson, the more senior physician, performs a more thorough history and exam with a broader differential in mind, and notes signs of biventricular failure that the resident missed. These findings increase her suspicion for viral myopericarditis, a diagnosis which becomes more likely with additional history that Andre has felt increasingly unwell since a viral syndrome 1 month ago. That the decision support system did not rank viral myopericarditis high on the potential list of diagnoses is less a failure of the diagnostic algorithm than a failure of the human component. Cognitive informatics emphasizes a balanced approach to how humans and machines work together. One could imagine circumventing the resident's premature diagnostic closure by instrumenting Andre's existence—surveilling his exposure to a virus 1 month ago, tracking his progressively worsening symptoms and elevated heart pressures, sensing his decreased gait speed and mobility—to diagnose his condition before he hit the ER. Aside from the technical challenges of achieving accurate diagnosis using such multi-modal time-varying data, the continuous collection of vast amounts of data from our daily lives presents a potentially grave cost in privacy. Data privacy is a core element of trust, as is, increasingly, transparency and fairness of the algorithms underlying computational decision support (Chap. 18). The remainder of this chapter discusses the main data and computational issues raised by Andre's use case.

Computational Landscape

There exists frequent confusion between **artificial intelligence (AI)** and **machine learning (ML)** and between ML and statistics. AI is the ability of a machine to perform tasks (and behave) like an intelligent being. AI encompasses a broad range of functions that lead a machine to "seem" intelligent, that we can break up into functions relating to data acquisition and processing, "thinking", and action in the real world. Data acquisition and processing include machine vision and image processing (e.g., detecting breast cancer in a mammogram, Chap. 12), speech recognition (e.g., dialog systems, Chap. 9), and NLP (e.g., extracting smoking status from EHR free text, Chap. 7). Thinking includes reasoning (as above), planning (e.g., surgical robot planning), and learning (Chaps. 5 and 6). Action in the real world includes image generation (e.g., embodied conversational agents, Chap. 9), speech generation (e.g., dialog systems, Chap. 9) and autonomous systems (e.g., robots that deliver meds).

As shown in Fig. 3.2, AI is a subset of computer science and ML is a subset of AI. Confusingly, ML also overlaps with statistics and data science. In fact, if ML is "computer algorithms that find and apply patterns in data," statistics does so too. Although ML typically is used on huge amounts of data, both ML and statistics are just alternative ways to understand and draw inferences out of data. Because ML

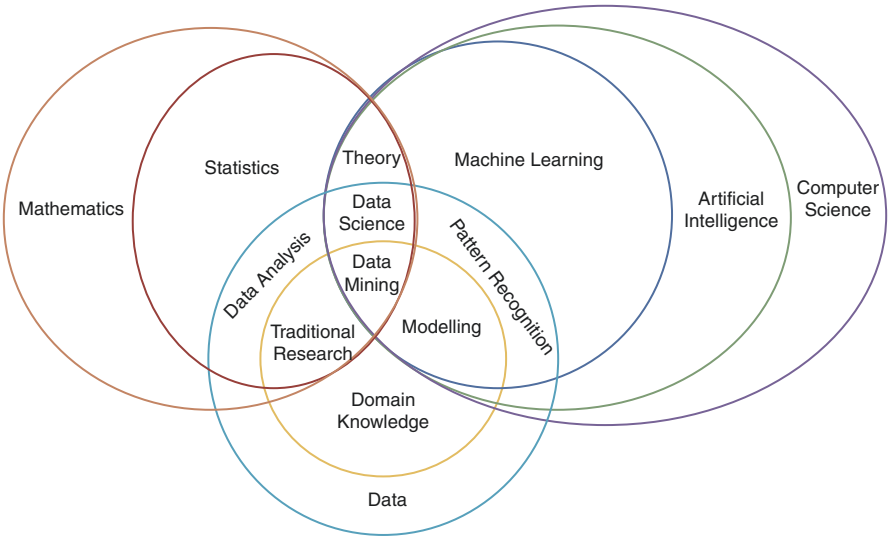

Fig. 3.2 Data Science Field/Term Diagram (adapted from Ryan J Urbanowicz, ryanurb@upenn.edu)

has commonalities with traditional analytics, ML is subject to the same pitfalls as traditional statistics, including bias, confounding, or inappropriate interpretation (Chap. 18). We can and should hold ML to the same expectations for scientific integrity as we do traditional analytics.

Knowledge Representation

Knowledge representation is the field of AI dedicated to representing information about the world in a form that a computer system can understand and use to solve complex tasks such as diagnosing a medical condition.

There are different approaches to data representation including symbolic, rule-based and graph-based formalisms. One of the most active areas of knowledge representation research are projects associated with the Semantic Web which seeks to add a layer of meaning on top of the internet. Rather than indexing web sites and pages via keywords, the Semantic Web creates large ontologies of concepts. An **ontology** is a set of concepts and categories in a subject area or domain that shows their properties and the relations between them. An example of an ontology in the biomedical domain is the Gene Ontology used to annotate genes.

A **rule-based** system has a knowledge base represented as a collection of "rules" that are typically expressed as "if-then" clauses. The set of rules forms the knowledge base that is applied to the current set of facts.

One of the earliest examples of such a system in the clinical domain was **MYCIN** [37], an early **backward chaining** expert system that used artificial intelligence to identify bacteria causing severe infections, such as bacteremia and meningitis, and to recommend antibiotics, with the dosage adjusted for a patient's body weight. Knowledge graphs are another method to model knowledge. A **knowledge graph** is a directed, labeled graph in which the labels have well-defined meanings. A directed labeled graph consists of nodes, edges (links), and labels. Anything can act as a node, for example genes, proteins, diagnoses. Edges between them can be relationships. This type of representation can be used for predicting and modeling different biological associations for instance drug-protein targets, gene-disease associations, protein-protein interactions, disease comorbidities. Knowledge-based systems are discussed in detail in Chap. 4 of this volume.

Machine Learning

Machine learning is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Machine learning approaches which are in further detail described in Chap. 6 can be characterized into **supervised** and **unsupervised** approaches (Fig. 3.3). **Clustering** algorithms that aim to group objects with similar attributes using measures of distance or similarity. For instance, one can cluster patients based on their clinical profiles and identify subgroups of patients that might be similar to each other. Unsupervised algorithms, or those that do not rely on ground truth, include k-means, hierarchical clustering, and expectation-maximization clustering using Gaussian mixture models. Classification is a task of identifying which category an observation belongs to. Some examples include classifying an email to the "spam" or "non-spam" category, or in the biomedical domain, assigning a diagnosis to a given patient based on observed characteristics of the patient. **Classification** algorithms, which often rely on training data, include random forest, decision trees, naive bayes and others and are supervised, which means that there is some data that is used with existing labels. These concepts are further explored in Chap. 6.

Deep learning techniques deserve special mention due to the importance these methods are gaining currently. Deep learning methods rely on neural networks, which were first proposed in the 1940s, in which layers of neuron-like nodes mimic

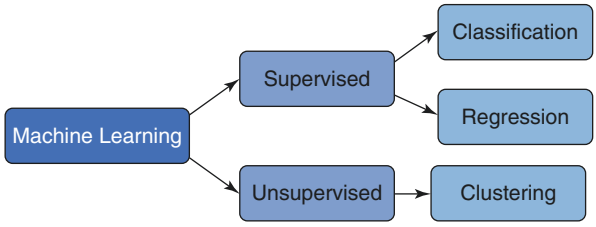

Fig. 3.3 Types of Machine Learning (ML) Approaches

how human brains analyze information. The underlying mechanisms of trained neural networks can be hard to disentangle, and thus, they have mainly been applied within biomedicine for image recognition. However, the ability to train a neural network on massive amounts of data has raised special interest in applying them elsewhere in the field of biomedicine, although the interpretability of these approaches is often a challenge (Chap. 8). These methods have been applied extensively to image analysis [38] and have been recently extended to other types of data including EHR [39] and genetic data [40].

Data Integration to Better Understand Medicine: Multimodal, Multi-Scale Models

The wealth and availability of public genomic, transcriptomic and other types of molecular data together with rich clinical phenotyping and computational integrative methods provide a powerful opportunity to improve human health by refining the current knowledge about disease therapeutics and diagnostics. There are different types of integrative models that can be applied to bring together diverse data [41]. As presented by Richie et al., meta-dimensional analysis can be divided into three categories: (1) **Concatenation-based integration** involves combining data sets from different data types at the raw or processed data level before modelling and analysis; (2) **Transformation-based integration** involves performing mapping or data transformation of the underlying data sets before analysis, and the modelling approach is applied at the level of transformed matrices; and (3) **Model-based integration** is the process of performing analysis on each data type independently, followed by integration of the resultant models to generate knowledge about the trait of interest.

The ideal scenario is when the different types of data are collected on the same individuals. In this case both concatenation and transformation-based integration can be applied. In our prior work, we examined patient heterogeneity in a lupus cohort for which we had rich clinical as well as molecular measurements such as genotyping and methylation to identify several clinical clusters of SLE patients and molecular pathways associated with those clusters [42]. However, there are also situations when the data is not collected on the same individuals and therefore, we must use a model-based integration approach to bring the datasets together using phenotype as the common ground. For instance, if the goal is to identify genetic, transcriptomic and proteomic associations with a certain disease of interest, data sets could be extracted from the public domain, where DNA sequence data may be available on some of the patient samples, microarray data from a different subset of patient samples, and proteomic data on yet another subset of patient samples. Model-based integration would allow the independent analysis of each of the modalities, followed by an integration of the top models from each data set to identify integrative consensus models.

By integrating data across measurement modalities as well as by integrating molecular measures with rich clinical phenotyping we can get a bit closer to achieving precision medicine by improving diagnostics and therapeutics.

Distributed/Networked Computing

The modern world is a networked world including more recent technologies for computing such as cloud computing and graphical processing units (GPUs—more in Chap. 2), where both data and computation are distributed across time, space, and jurisdictions. A patient's EHR data may reside in several places: her current primary doctor's health system, the health systems of her previous care providers and of emergency visits, and third party telehealth companies for the occasional urgent care consult. Her genetic sequence, cancer genotype, and various wearable data such as Apple Watch or FitBit are likewise held and computed on in siloed and proprietary systems, each subject to access policies and terms of use that are often opaque to both the patient herself and to third parties. As messy as this all is, as discussed in the section on "Types of Data Relevant to Biomedicine", there is no such thing as "health" and "non-health" data. Because the value of data is in its aggregation, a challenge is how to bring together multiple sources of data for any given query to enable multiple types of computation.

In the traditional approach, data is brought to the query. That is, if a data requester wants to run a query, the requester obtains a copy of the data, installs it on his/her own machine and runs the query on the data that has been brought in. Because the data requester now holds a copy of the data, the original data holder has effectively lost control over its access. Moreover, if the datasets are very large, as is the case for many imaging, genomic, sensor, and real-world data studies, the data requester may not have sufficient storage and compute capacity. Thus, this approach is not compatible with any need for controlled access (which includes most cases of sharing patient data) nor for sharing large datasets.

The converse approach is bringing the query to the data. The data requester submits the query to the machine where the data resides, the query is run on that remote machine, and the results are returned back to the requester. Queries can, of course, be complex computations and analyses, not just simple search and retrieval queries. In this model, data holders retain control of the data and the requester does not ever have a copy of or control of the data.

Data Federation Models

This basic idea of bringing the query to the data can be implemented through different configurations of databases and query servers, each with their own benefits and challenges. In the simplest **Local Data Store** model, every data holder hosts its own data on its own server. External data requesters establish user accounts on that

server under some access control model. The requester then has access to view the data and to analyze it, but not to download a copy of the data to the requester's own machine. However, this model is infeasible for widespread data sharing because data requesters wishing to query multiple databases must establish multiple user accounts and navigate multiple access policies and procedures and have no ability to combine data for aggregate analysis.

In the **One Single Centralized Datastore** model, data from multiple sources are aggregated into one "data warehouse." An example is the University of California's Health Data Warehouse that aggregates data from over 15 million patients seen at the five medical campuses of the University of California [43]. Another example is N3C, aggregating EHR data on 1.9 million COVID patients from 34 medical centers across the US into a single portal for secure data access and analysis [44]. This model benefits from economies of scale, and data requesters need to submit their queries to only one database under a uniform data access policy. However, this model still does not allow aggregation of data across data warehouses. The silo is just a bigger silo.

The **federated query model** combines the bring-the-query-to-the-data approach with federated databases. Databases are federated when independent geographically dispersed databases are networked in such a way that they can respond to queries as if all the data were in a single virtual database. Thus, data requesters can submit a query to a federated query service and have that query be routed to all databases participating in that federation. Data holders maintain full control of their data, and neither the data requester nor the query service provider ever has direct access to the data.

Federation technology has progressed substantially in recent years. An example is the R2D2 initiative with a federated network of 12 health systems comprising 202 hospitals contributing COVID-related EHR data on 45 million patients [45]. In contrast to the N3C approach described above, data never leave the 12 health systems, which act as nodes on the network making their patient data available in a common data model. Queries and computations are submitted via a centralized service that then federates computation such as averages, regressions, and machine learning models to individual nodes on the network.

Interoperability

Whenever data is brought together for query and computation, whether in the centralized warehouse or federated model, the data must be interoperable. **Interoperability** is the ability of computer systems or software to exchange *and make use of* information; it is not enough to send data that is unintelligible to the recipient. Interoperability therefore includes both syntactic and semantic interoperability, which are enabled by the use of data interoperability standards. **Syntactic interoperability** refers to the format and ordering of what is exchanged, analogous to the grammar of an English sentence for exchanging ideas between humans. Examples of primarily syntactic standards include data exchange standards such as

HTML and, in health care, HL7 FHIR and DICOM for representing data in transit. **Semantic interoperability** refers to the meaning of what is exchanged, analogous to the words and their dictionary meaning in an English sentence. Semantic standards in healthcare include common terminologies such as SNOMED and LOINC. One needs both syntactic and semantic standards to enable full interoperability. A sentence using English words and German grammar is not interoperable between humans (Box 3.3).

Box 3.3 Interoperability
Interoperability *is the ability of computer systems or software to exchange and make use of information.* ***Syntactic interoperability*** *refers to the format and ordering of what is exchanged.* ***Semantic interoperability*** *refers to the meaning of what is exchanged.*

Data that is to be aggregated also need to share a common data model at rest. The University of California Health Data Warehouse, N3C warehouse, and the R2D2 federated network all use the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) [46]. This model was designed for cross-institutional queries of EHR data for quality improvement and research purposes, and binds data to a mandatory clinical vocabulary (OMOP Standardized Vocabularies) [47] that is based on SNOMED, LOINC, RxNORM and others. Note the same OMOP data model and associated vocabularies can be used for centralized or federated approaches and is fit-for-purpose for a wide range of EHR data interoperability use cases. Common data models and data exchange protocols must be defined and agreed upon across all contributors to data sharing and adopted uniformly by each contributor or federation node.

Computational Aspects of Privacy

Chapter 18 reviews the broader issues of Ethics, including Privacy. To understand the computational aspects of privacy, we need to distinguish privacy and security. **Privacy** is a concept that applies to people, rather than documents, in which there is a presumed right to protect that individual from unauthorized divulging of personal data of any kind. **Security** is the process of protecting information from destruction or misuse, including both physical and computer-based mechanisms. Security falls under IT. Privacy is when you are assured and protected from a company holding your geolocation data selling it without your knowledge or approval. Security is when no hacker can get into that company's systems to access or corrupt your geolocation data. You can have 100% security and no privacy; if you have no security, you also have no privacy.

Privacy is best protected by a combination of technical and legal means. Technically, the objective is to minimize the risk that an adversary can associate or re-identify your personal data with you. However, there is no guarantee of absolute protection against re-identification. Data—EHR, geolocation, fitness data—can be subjected to **de-identification** or **anonymization** to increase privacy. De-identified data is data that has identifying personal data such as names and birthdates removed or perturbed in such a way as to be non-identifying (Chap. 18). Anonymized data has identifying personal data removed or perturbed and the key linking a data record to a particular person is destroyed such that the data becomes anonymous. In truth, with sufficient external data, de-identified and even anonymized data can be re-identified. Thus, legal mechanisms such as data use agreements (DUA) are needed to supplement technical privacy protection.

The challenges of privacy protection are magnified when data need to be aggregated. The more data there is about an individual, the greater the risk the data will uniquely match an individual leading to re-identification.

The risk is further magnified with federated data sharing. Mechanisms such as differential privacy [48] are used where queries are federated over perturbed data and the answers are then operated on to "subtract" out the perturbations to arrive at the real answer without increasing the risk of re-identification. Another approach is synthetic data [49], where a synthetic dataset is created that matches the distributional properties of the original data set. In this way, computations can occur on the synthetic dataset with some provable level of accuracy to the original dataset. The details of computation for privacy are outside the scope of this book but are closely tied to the ability to safely aggregate and reuse large amounts of data for machine learning. The point to know is that the "old" way of privacy protection under HIPAA "safe harbor" [50]—removing a specified list of 18 identifying data elements—is increasingly insufficient for modern-day data sharing and computation.

Trends and Future Challenges

Chapter 19 anticipates the future of AI in medicine and healthcare. Here, we review trends and open challenges affecting the general use of data and computation for biomedicine.

Ground Truth

The availability of extensive molecular and clinical data provides an incredible opportunity to apply predictive modeling and ML techniques to improve diagnostics and therapeutics. However, ML models need rich and accurate training data, including labelling of ground truth (e.g., which patients have the disease that the

ML is trying to predict). Many datasets, especially large public datasets, are poorly labelled and are also heterogeneous and not well annotated, making them difficult to aggregate and use for ML. Annotation and labelling are difficult and time-consuming tasks. For example, labelling clinical and imaging data with ground truth labels of diagnosis requires expert time and costs. The limited availability of labelled data can be somewhat overcome by the sheer amount of data, but this bottleneck is important to recognize. New semi-supervised approaches are emerging that rely on small amounts of labelled data to predict missing labels for larger datasets, but these approaches risk perpetuating and amplifying biases and mis-labelling in the smaller set. The availability of accurate and unbiased labelled training data for ML will be an ongoing challenge.

Open Science and Mechanisms for Open data

Scientific culture is increasingly embracing open approaches to data sharing and reuse, adhering to FAIR (Findable, Accessible, Interoperable, Reusable) principles [51]. **Findability** requires indexing and shared metadata and persistent Digital Object Identifiers (DOIs) such as from DataCite or other services that span disciplines. **Accessibility** brings up data rights, ownership, access policies, and fair (as in just) credit for data sharing—all of which are wide open issues. We discussed **Interoperability** above.

Reusability needs to be distinguished between reusability by humans or reusability by computers. Human reusability is a lower bar. Data and metadata need to be findable and sufficiently interpretable by humans to facilitate additional data cleaning, alignment, and harmonization to achieve the aggregation purpose. In contrast, automated reusability by computers requires much more stringent adherence to compatible syntactic and semantic standards for all the data. This upfront work is challenging for data mapping but also governance reasons. People need to decide on common data elements, which necessitate agreement on potentially controversial scientific issues. For example, the N3C Consortium agreed on specific definitions of variables that all sites have to map their data into [52]. N3C also had to address privacy concerns for human data, as data reusability must take place under fair and just conditions that limit the risk of re-identification. Thus, N3C also defined three levels of access to N3C data in their secure enclave: a limited data set that can only be accessed with IRB approval, a de-identified data set that can be accessed without IRB approval, and a synthetic data set that requires only an N3C account and DUA [53]. Perhaps because of the additional challenges of protecting privacy, the ethos of open science and open data has a stronger hold in the life sciences than in clinical research and care.

As data, information and knowledge are shared, re-purposed, combined, and distributed in a networked world, the **provenance** of each component must be auditable lest errors and biases become compounded to an extent that threatens the integrity of computed inferences and decision support. Infrastructures for managing metadata and provenance are currently woefully inadequate. The FHIR [54] and

Open mHealth [55] data exchange standards model provenance (e.g., who measured a blood pressure reading in the clinic, what sensor model did the sleep duration come from) but these attributes are not consistently captured or described. The need for detailed provenance is critical for scientific reproducibility and is especially important for longitudinal studies of data that may drift over time. For example, the NIH's 4-year RECOVER initiative to study Post-Acute Sequelae of SARS-CoV-2 (PASC, aka "long COVID") will be collecting real-world, survey, and sensor data whose definition, collection, and post-processing methods are likely to change as more is known about PASC. Without a clear trace of data transformations and other provenance, the scientific value of the consortium data will diminish over time. Provenance architectures, managing the risks of re-identification, and mechanisms for tracking and assigning data sharing credit are two major open challenges.

Data as a Public Good

The ultimate value of data and computation to society rests on the willingness of the intended users to accept the outputs. How we collect, describe, and share data and how we construct our computational systems can earn the trust of users—or not. As discussed in Chap. 18, trustworthiness must be designed into data and computation from the outset and cannot be left as an afterthought. Lack of trust is corrosive and impedes data fluidity and data aggregation, which decreases the overall value of computation by reducing the amount and representativeness of the underlying data.

One of the central pillars of trustworthiness involves protecting the privacy of individuals. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) [50] governs health data privacy by regulating healthcare organizations ("covered entities") on when they can use and disclose individuals' health information. This approach implicitly sets healthcare organizations as the principal custodians of health data, thus giving such organizations outsize control (and responsibility) over the trust fabric for the use of computers in health care. The European Union, in contrast, takes a person-centered rather than an organization-centered approach. The General Data Protection Regulation (GDPR) [56] explicitly places the individual in control of the use and disclosure of their own data, and defines a more expansive framing of data protection to include not only privacy but also appropriately scoped data requests, transparency, and fairness. As demonstrated by the SARS-CoV-2 pandemic, however, data also serve as a public good to inform public policies, drive machine learning models, or demonstrate the efficacy of pharmaceutical and non-pharmaceutical interventions. While care must be taken to repurpose data originally collected for individual care, a justice-based model for data sharing [57] is emerging that focuses on fostering public trust in uses of such data for the public good with attention to the needs of vulnerable populations and eliminating health disparities. Data sharing that prioritizes public interest as well as personal privacy promotes optimal data use for society. Over time, the technical architecture of data, data sharing, and computation will morph to drive and align

with society's evolving relationship to data, with deep implications for the future of cognitive informatics.

Questions for Discussion

- What are some existing clinical data resources and standards that allow for data analysis and integration?
- What molecular databases exist now that can be leveraged for biomedical research?
- How can supervised and unsupervised machine learning approaches complement traditional evidence-based medicine approaches?
- What is data federation and in what ways can it be achieved?
- What kind of data can be integrated to impact clinical decision making and care?
- What are key considerations in ensuring trustworthy data and computation?

Further Reading

Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015 Feb;16 (2):85–97. doi: 10.1038/nrg3868. Epub 2015 Jan 13. PMID: 25582081.

- This is a comprehensive review paper on integrative approaches. The authors explore meta-dimensional and multi-staged analyses — in the context of better understanding the role of genetics and genomics in complex outcomes. However the aforementioned approaches can also be leveraged for integrating other types of data.

Sim I. Mobile Devices and Health. N Engl J Med 2019; 381:956–968.

- Comprehensive review of leveraging mobile devices in health. This article discusses sensors, digital biomarkers, digital therapeutics and diagnostics, and the integration of mobile health into frontline clinical care. It concludes with open questions on the ethics, validation, and regulation of mobile health and the prevailing market forces that are shaping the growth of this technology sector.

Straus S, Glasziou P, Richardson WS, Haynes RB. (2018) *Evidence-Based Medicine: How to Practice and Teach It.* Elsevier. ISBN: 9780702062964.

- A comprehensive description of evidence-based medicine geared towards practicing clinicians. It reviews EBM approaches for the major types of clinical questions (therapy, diagnosis and screening, prognosis) and includes tools and calculators for teaching and applying EBM in practice.

Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A. Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, Victor Robles, Machine learning in bioinformatics, *Briefings in Bioinformatics*, Volume 7, Issue 1, March 2006, Pages 86–112, https://doi.org/10.1093/bib/bbk007.

- This is a comprehensive review of machine learning in bioinformatics. The authors present a number of modelling methods, such as supervised classification, clustering and probabilistic graphical models for knowledge discovery, as well as deterministic and stochastic heuristics for optimization. They present applications in genomics, proteomics, systems biology, evolution and text mining however the methodology is applicable to other types of data including clinical.

Alyass, A., Turcotte, M. & Meyre, D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics 8, 33 (2015). https://doi.org/10.1186/s12920-015-0108-y.

- While there are incredible opportunities with the recent advances in high throughput technologies allowing for leveraging and integrating large datasets to achieve more precise modeling of human disease, there are also challenges that need to be recognized. Several bottlenecks include generation of cost-effective high-throughput data; hybrid education and multidisciplinary teams; data storage and processing; data integration and interpretation; and individual and global economic relevance. This article discusses challenges and opportunities in personalized medicine using big data.

References

1. The world's most valuable resource is no longer oil, but data. The Economist [Internet]. 2017 May 6 [cited 2021 June 13]. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.
2. Are data more like oil or sunlight? The Economist [Internet]. 2020 Feb 20 [cited 2021 June 13]. https://www.economist.com/special-report/2020/02/20/are-data-more-like-oil-or-sunlight.
3. Ackoff RL. From data to wisdom. J Appl Syst Anal. 1989;16:3–9.
4. EPISTEMOLOGY | Definition of EPISTEMOLOGY by Oxford Dictionary on Lexico.com also meaning of EPISTEMOLOGY [Internet]. Lexico Dictionaries | English. [cited 2021 June 13]. https://www.lexico.com/en/definition/epistemology.
5. Beck AH. STUDENTJAMA. The Flexner report and the standardization of American medical education. JAMA. 2004;291(17):2139–40.
6. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA. 1992;268(17):2420–5.
7. Clarke B, Gillies D, Illari P, Russo F, Williamson J. The evidence that evidence-based medicine omits. Prev Med. 2013;57(6):745–7.
8. Oxford Centre for Evidence-Based Medicine: Levels of Evidence (March 2009) — Centre for Evidence-Based Medicine (CEBM), University of Oxford [Internet]. [cited 2021 July 23]. https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009.
9. RECOVERY Collaborative Group, Horby P, Lim WS, Emberson JR, Mafham M, Bell JL, Linsell L, Staplin N, Brightling C, Ustianowski A, Elmahi E, Prudon B, Green C, Felton T, Chadwick D, Rege K, Fegan C, Chappell LC, Faust SN, Jaki T, Jeffery K, Montgomery A,

Rowan K, Juszczak E, Baillie JK, Haynes R, Landray MJ. Dexamethasone in hospitalized patients with Covid-19. N Engl J Med. 2021;384(8):693–704.
10. Sim I. Two ways of knowing: big data and evidence-based medicine. Ann Intern Med. 2016;164(8):562–3.
11. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18.
12. Abraham A, Le B, Kosti I, Straub P, Velez-Edwards DR, Davis LK, et al. Dense phenotyping from electronic health records enables machine-learning-based prediction of preterm birth. medRxiv. 2020;2020.07.15.20154864.
13. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 1.0). PhysioNet. 2021. https://doi.org/10.13026/s6n6-xd98.
14. Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, et al. ImmPort: disseminating data to the public for the future of immunology. Immunol Res. 2014;58(2–3):234–9.
15. Vivli - Center for Global Clinical Research Data [Internet]. [cited 2021 July 22]. https://vivli.org/.
16. Clough E, Barrett T. The gene expression omnibus database. Methods Mol Biol. 2016;1418:93–110.
17. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007;39(10):1181–6.
18. Tomczak K, Czerwińska P, Wiznerowicz M. The cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68–77.
19. Sirota M, Thomas CG, Liu R, Zuhl M, Banerjee P, Wong RJ, et al. Enabling precision medicine in neonatology, an integrated repository for preterm birth research. Sci Data. 2018;5:180219.
20. Admin O. CalEnviroScreen [Internet]. OEHHA. 2014 [cited 2021 July 22]. https://oehha.ca.gov/calenviroscreen.
21. Office of Statewide Health Planning and Development [Internet]. OSHPD. [cited 2021 July 22]. https://oshpd.ca.gov/.
22. Wang A, Gerona RR, Schwartz JM, Lin T, Sirota M, Morello-Frosch R, et al. A suspect screening method for characterizing multiple chemical exposures among a demographically diverse population of pregnant women in San Francisco. Environ Health Perspect. 2018;126(7):077009.
23. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation - PubMed [Internet]. [cited 2021 July 20]. https://pubmed.ncbi.nlm.nih.gov/31722151/.
24. Wesley DB, Blumenthal J, Shah S, Littlejohn R, Pruitt Z, Dixit R, et al. A novel application of SMART on FHIR architecture for interoperable and scalable integration of patient-reported outcome data with electronic health records. J Am Med Inform Assoc. 2021;28(10):2220–5. https://doi.org/10.1093/jamia/ocab110.
25. Tong HL, Quiroz JC, Kocaballi AB, Fat SCM, Dao KP, Gehringer H, et al. Personalized mobile technologies for lifestyle behavior change: a systematic review, meta-analysis, and meta-regression. Prev Med. 2021;148:106532.
26. Milne-Ives M, Lam C, De Cock C, Van Velthoven MH, Meinert E. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. JMIR Mhealth Uhealth. 2020;8(3):e17046.
27. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol. 2008;4:1–32.
28. Catlett CE, Beckman PH, Sankaran R, Galvin KK. Array of things: a scientific research instrument in the public way: platform design and early lessons learned. In: Proceedings of the 2nd International Workshop on Science of Smart City Operations and Platforms Engineering - SCOPE '17. Pittsburgh, PA: ACM Press; 2017. http://dl.acm.org/citation.cfm?doid=3063386.3063771.
29. Neelon B, Mutiso F, Mueller NT, Pearce JL, Benjamin-Neelon SE. Associations between governor political affiliation and COVID-19 cases, deaths, and testing in the U.S. Am J Prev Med. 2021;61(1):115–9.
30. Peirce CS. In: Buchler J, editor. Philosophical writings of Peirce. New York: Dover; 1955.

31. Harman G. The inference to the best explanation. Philos Rev. 1965;74:88–95.
32. Bureau UC. Health Insurance Coverage in the United States: 2019 [Internet]. The United States Census Bureau. [cited 2021 July 22]. https://www.census.gov/library/publications/2020/demo/p60-271.html.
33. Health Coverage Affordability Crisis 2020 Biennial Survey | Commonwealth Fund [Internet]. [cited 2021 July 22]. https://www.commonwealthfund.org/publications/issue-briefs/2020/aug/looming-crisis-health-coverage-2020-biennial.
34. Sisodia RC, Rodriguez JA, Sequist TD. Digital disparities: lessons learned from a patient reported outcomes program during the COVID-19 pandemic. J Am Med Inform Assoc. 2021;28(10):2265–8. https://doi.org/10.1093/jamia/ocab138.
35. An early warning scoring system for detecting developing critical illness – ScienceOpen [Internet]. [cited 2021 July 22]. https://www.scienceopen.com/document?vid=28251d22-8476-40a6-916d-1a34796816e4.
36. McSherry D. Avoiding premature closure in sequential diagnosis. Artif Intell Med. 1997;10(3):269–83.
37. Shortliffe EH. Mycin: a knowledge-based computer program applied to infectious diseases. Proc Annu Symp Comput Appl Med Care. 1977;5:66–9.
38. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.
39. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep. 2016;6(1):26094.
40. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20(7):389–403.
41. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16(2):85–97.
42. Lanata CM, Paranjpe I, Nititham J, Taylor KE, Gianfrancesco M, Paranjpe M, et al. A phenotypic and genomics approach in a multi-ethnic cohort to subtype systemic lupus erythematosus. Nat Commun. 2019;10(1):3902.
43. Center for Data-driven Insights and Innovations (CDI2) | UCOP [Internet]. [cited 2021 July 22]. https://www.ucop.edu/uc-health/functions/center-for-data-driven-insights-and-innovations-cdi2.html.
44. Bennett TD, Moffitt RA, Hajagos JG, Amor B, Anand A, Bissell MM, et al. The National COVID Cohort Collaborative: clinical characterization and early severity prediction. medRxiv. 2021;2021.01.12.21249511.
45. Kim J, Neumann L, Paul P, Day ME, Aratow M, Bell DS, et al. Privacy-protecting, reliable response data discovery using COVID-19 patient observations. J Am Med Inform Assoc. 2021;28(8):1765–76. https://doi.org/10.1093/jamia/ocab054.
46. Chapter 4 The common data model | The book of OHDSI [Internet]. [cited 2021 July 22]. https://ohdsi.github.io/TheBookOfOhdsi/.
47. Chapter 5 Standardized vocabularies | The book of OHDSI [Internet]. [cited 2021 July 22]. https://ohdsi.github.io/TheBookOfOhdsi/.
48. Dwork C, Roth A. The algorithmic foundations of differential privacy. Found trends®. Theor Comput Sci. 2014;9(3–4):211–407.
49. El Emam K, Mosquera L. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. Sebastopol, CA: O'Reilly Media, Inc; 2020.
50. Office for Civil Rights. Summary of the HIPAA privacy rule [Internet]. HHS.gov. 2008 [cited 2021 July 22]. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.
51. Wilkinson MD, Dumontier M, IJJ A, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018.
52. COVID-19 Clinical Data Warehouse Data Dictionary Based on OMOP Common Data Model Specifications Version 5.3. :22.

53. N3C data overview [Internet]. National Center for Advancing Translational Sciences. 2020 [cited 2021 July 22]. https://ncats.nih.gov/n3c/about/data-overview.
54. Overview - FHIR v4.0.1 [Internet]. [cited 2021 July 22]. https://www.hl7.org/fhir/overview.html.
55. Open mHealth [Internet]. GitHub. [cited 2021 Jul 22]. https://github.com/openmhealth.
56. EUR-Lex - 32016R0679 - EN - EUR-Lex [Internet]. [cited 2021 July 22]. https://eur-lex.europa.eu/eli/reg/2016/679/oj.
57. University of California. President's Ad Hoc Task Force on Health Data Governance [Internet]. 2018 [cited 2021 July 22]. https://www.ucop.edu/uc-health/reports-resources/health-data-governance-task-force-report.pdf.