Published in final edited form as:

J Pediatr. 2019 May ; 208: 12–22. doi:10.1016/j.jpeds.2018.12.041.

Data Science for Child Health

Tellen D. Bennett, MD, MS^{1,2,3,4,5}, Tiffany J. Callahan, MPH⁵, James A. Feinstein, MD, MPH^{1,4}, Debashis Ghosh, PhD^{2,3,5}, Saquib A. Lakhani, MD⁶, Michael C. Spaeder, MD, MS⁷, Stanley J. Szefer, MD^{1,4}, and Michael G. Kahn, MD, PhD^{1,5}

¹Department of Pediatrics, University of Colorado School of Medicine

²CU Data Science to Patient Value (D2V), University of Colorado School of Medicine

³Biostatistics and Informatics, Colorado School of Public Health

⁴Adult and Child Consortium for Outcomes Research and Delivery Science (ACCORDS), University of Colorado School of Medicine and Children's Hospital Colorado

⁵Computational Bioscience Program, University of Colorado Denver Anschutz Medical Campus, Aurora, CO

⁶Pediatric Genomics Discovery Program, Department of Pediatrics, Yale University School of Medicine, New Haven, CT

⁷Pediatric Critical Care, University of Virginia School of Medicine, Charlottesville, VA

Data science has revolutionized industry and academic fields including marketing,¹ astronomy,² and computer vision.³ It has not yet impacted medicine and biomedical research to the same degree. However, many observers^{4–6} believe that data science will improve the ability of health care systems to deliver personalized medicine,⁷ population health,⁸ and public health.⁹

The US National Institutes of Health (NIH) Strategic Plan for Data Science defines it as “the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data.”¹⁰ One interpretation of this definition is that data science is possible because of relatively cheap data storage and computing, efficient algorithms and computational tools for complex tasks such as optimizing neural network models and analyzing whole exome sequence (WES) data, and the availability of enormous amounts of data. This includes biological (eg, genomic, proteomic, and metabolomic), clinical (eg, laboratory results, digital images, text notes, and monitor signals), lifestyle (mobile sensor output, interactions with applications), and other data with which to train and evaluate models.

Reprint requests: Tellen D. Bennett, MD, MS, Associate Professor, Pediatric Critical Care, University of Colorado School of Medicine, Children's Hospital Colorado, Biostatistics and Informatics, Colorado School of Public Health (Secondary), Co-Director, Analytics Core, CU Data Science to Patient Value, Associate Director, Informatics Core, Colorado Clinical and Translational Sciences Institute (CCTSI), 13199 E. Montview Blvd, Suite 300, Mail Stop F443, Aurora, CO 80045. tell.bennett@ucdenver.edu.

The objectives of this review were to identify unique aspects of pediatrics that are relevant to the potential impact of data science on child health and provide examples of how common data science tasks are being used to facilitate clinical and biomedical child health research.

Why Data Science for Child Health?

At least 3 unique features of pediatrics are important when considering the potential impact of data science. First, most children interact with the health care system infrequently and primarily for well-child visits. Compared with subspecialty and inpatient care, infrequent outpatient pediatric primary care generates relatively little data from which to glean new and actionable knowledge, no matter how sophisticated the available algorithms. By extension, diseases that affect children are relatively rare and often heterogeneous in presentation and phenotype.¹¹ Because most children are well, medical centers and health systems tend to care for fewer sick children than sick adults. Therefore, condition-specific pediatric single-center studies are less likely to have enough patients or samples to achieve scientific goals. To assemble sufficient data to train machine learning models (which frequently require large training datasets), for example, large multicenter studies may be necessary. Because of this, research networks and data sharing are even more important for child health research than for adult health research.

A second unique feature of pediatrics is that most congenital conditions present to clinical care during childhood, often during infancy. These congenital conditions may be quite rare and may demonstrate locus heterogeneity; similar phenotypes may be the result of mutations in different genes. Because these conditions present in childhood, the potential for genomics, proteomics, transcriptomics, metabolomics, and so on (collectively, the “omics”) to identify novel diseases and generate hypotheses about potential treatments is greater in pediatrics than in adult medicine. Recently developed data science methods and tools have rapidly increased the clinical and research impact of omics technologies.

Finally, temporality is more important in child health research than in other fields. Because of child development and puberty, it is critical to know a child’s age at a more granular resolution than is typically necessary in adult medicine. Child development over time also powerfully influences heterogeneity in presentation and phenotype of pediatric diseases. Every pediatrician is trained to recognize that, for example, asthma and leukemia present and are treated differently in infants compared with school-age children. Normal values for laboratory (eg, creatinine concentration) and vital sign (eg, heart rate, blood pressure) data often included in predictive models vary widely across the pediatric age range. The accurate timing of events is also important for studying seasonal infectious diseases. Respiratory and gastrointestinal viruses cause many well children to require health care and worsen many chronically ill children.¹² The same viruses sicken many adults, but they cause adults to require emergency and inpatient care less often than they do young children.

Issues with Data: Availability, Volume, and Complexity

Clinical, Biological, and Lifestyle Data Sources

The most important raw materials for data science are, of course, data. Common clinical data sources include the structured contents of electronic health records (EHRs) including vital signs, laboratory results, and medication dose and timing. Structured data are in standardized formats that computers can understand (eg, numerical heart rate, birth date in a specified date format, medications using accepted drug codes). EHRs also contain large amounts of unstructured text (eg, progress and operative notes, pathology and radiology reports). Ancillary systems connected to EHRs such as waveform archives and picture archiving and communication systems contain huge amounts of signal (eg, from monitors) and raw digital image data. Natural language processing tools, signal processing algorithms, and computer vision techniques can extract clinically actionable information from text, signals, and images, respectively. Until recently, such tools have not been widely available.

Common sources of biological data are genomic, proteomic, and metabolomic platforms. Data from these platforms are often too big to be analyzed using commodity computer systems and traditional statistical methods. Social networks, mobile devices, and self-tracking tools¹³ are also rich data sources.

The NIH Strategic Plan: FAIR Principles, Data Sharing, and Deidentification

The NIH Strategic Plan for Data Science highlights that many current data resources are siloed in such a way that they are not easily findable, shared, or interoperable.¹⁰ The FAIR (Findable, Accessible, Interoperable, and Re-usable)¹⁴ principles for data and data products (eg, software, models, etc) are a consistent point of emphasis for the NIH and other funders.¹⁵ For data to be findable, they must be shared. However, data privacy and security concerns can limit sharing. Temporal information, for example, is necessary to answer many important questions in child health, but dates more specific than a calendar year are protected health information. Deidentification (the removal of protected health information) is often a requirement for data sharing. For example, Medical Information Mart for Intensive Care III (MIMIC) is a publicly available and widely used adult critical care data resource that includes patient ages and dates that are randomly shifted into the future.¹⁶ Patients more than 89 years old are relatively rare and their ages are masked in the database. Deidentification would be a major challenge for an effort to share an analogous pediatric data source. Young infants require medical care more often than older children do, but date shifting of even a few days or weeks might destroy important temporal relationships. Completely masking ages in the database would decrease the database's value because investigators could not then accurately query for appropriate cohorts. A database with dates only as specific as the calendar year would be considered de-identified by the Health Insurance Portability and Accountability Act Privacy Rule Safe Harbor definition, but would have limited usefulness for child health research.

De-identification without losing temporal information by calculating intervals between, for example, the date of hospital admission and many potential events of interest (eg, intensive care unit [ICU] admission, a surgical procedure, mechanical ventilation, doses of particular

medications, etc) requires programmer effort and time at each site that contributes data. This may be a barrier to multisite data aggregation to support the application of data science methods.

Pediatric Research Networks

Despite these challenges, several successful pediatric-focused research networks exist. These include a National Pediatric Learning Health System (PEDSnet),¹⁷ the Pediatric Emergency Care Applied Research Network,¹⁸ the Collaborative Pediatric Clinical Care Research Network,¹⁹ the Pediatric Research in Inpatient Settings, and the Pediatric Research in Office Settings.²⁰ Two systems supported by the NIH Clinical and Translational Science Award program (Accrual to Clinical Trials and Informatics for Integrating Biology & the Bedside [i2b2]²¹) and a newer private initiative (TriNetX) will likely be important for multisite data sharing initiatives in the near future.

In the absence of a trusted third party such as a network data coordinating center to manage the sharing of data, probabilistic record linkage²² may be required to connect data sources that contain information about the same patients but do not share a unique identifier. Pediatric examples include linkages to support studies of children with cancer,²³ congenital heart disease (CHD),²⁴ and traumatic brain injury.²⁵

Data Standardization

Well-structured and standardized data facilitate data science. Clinical care and research databases store a vast array of data elements as discrete values (measurements, categories), text documents (plain text, PDF), or binary files such as static radiology, pathology, or dermatology images. Specialized systems are required to handle massive datasets such as genomic sequences, time-based physiologic monitoring streams, or real-time imaging modalities.^{26,27} Storing and then extracting data for large-scale data analytics involves understanding 2 features of the dataset or database: where relevant data elements are stored and how those data elements are coded. Data standards, terminologies, and data models support data-intensive research by structuring the coding (standards, terminologies) and location (data models) of the data.^{28,29} Common data elements³⁰ serve the same function for prospectively collected data.

For example, 1 database may contain all forms of diagnoses (admitting, discharge, billing, preoperative, postoperative, pathology, etc) in a single location (table, file). The various types of diagnoses could be distinguished using a diagnosis-type field. Another database may store each diagnosis type in separate tables (files). Differences in how a database are structured, called a data model, can have a significant impact on what clinical features can be stored and how difficult it may be to get data out for analytics.³¹

In addition to differences in storage location, values for diagnoses may be stored as free text strings or in one of many coding systems, such as the *International Classification of Diseases* (ICD) system, including the ICD-9-CM (historical) and the ICD-10-CM; Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT), or an EHR vendor's proprietary system. Because of the vast breadth of medical concepts, many overlapping and sometime inconsistent coding systems exist. The Unified Medical Language

System, a central repository of many US-specific and international medical coding systems maintained by the National Library of Medicine, currently contains 207 different terminologies.^{32,33} Unfortunately, pediatrics is poorly represented in most existing terminologies. Efforts are ongoing by pediatric specialty societies to address this deficiency.³⁴

Data Quality and Common Data Models

Another significant challenge for data science using data derived from EHR systems are concerns about data quality, especially missing or inaccurate data.^{35–37} Some data quality problems can arise because different EHRs store data in different formats. Despite decades of technical work and, more recently, US government mandates designed to reduce data interoperability barriers, exchanging and combining clinical data across multiple organizations remains a difficult task.^{38,39}

One widely used solution to combining data across different data sources is using a common data model (CDM).³¹ A CDM provides detailed instructions of how clinical data are structured (the data model); specific definitions and common codes to be used as valid values for each variable; how to handle missing, unusual, or unrealistic data values; and instructions for how to handle unusual edge cases that may appear in rare situations.⁴⁰ The key feature of a CDM is the word “common”; a CDM attempts to replace the large heterogeneity that exists in operational systems (EHRs, billing systems, ancillary systems) with a single data model structure, a set of valid codes, and a set of conventions for transforming data from its original format into the CDM. Although many CDMs exist, 4 currently dominate large-scale clinical research data-sharing efforts in the United States: i2b2,²¹ the virtual data warehouse CDM,⁴¹ the Observational Medical Outcomes Partnership CDM,⁴² and the Patient Centered Outcomes Research Network CDM.⁴³ For example, the multi-institutional PEDSnet CDM⁴² is based on the Observational Medical Outcomes Partnership.

Although they are less visible and exciting than cutting-edge modeling techniques, data standards, terminologies, common data elements, and data models are critical to the success of data science.

Using the Data

Machine Learning

Machine learning is not a necessary component of data science, but its approaches, models, and algorithms have lately received significant scientific and public recognition because of impressive performances automating and learning human tasks and games from scratch.^{44,45} Machine learning is a specialty within computer science that focuses on computers learning (eg, finding the optimal model or shortest path) from data without being explicitly programmed. Machine learning methods are currently being applied to a wide array of data types and questions in biomedical research. Because machine learning models can be flexible and very accurate, they have generated excitement about their potential application to prediction (prognostic) and classification (diagnostic) tasks in medicine.

Machine learning generally takes 2 forms, supervised and unsupervised learning (Figure). The goal of supervised learning is to predict an outcome (eg, disease yes/no, blood pressure). Unsupervised learning searches for previously unknown but quantitatively similar clusters or groups in a larger dataset. A more comprehensive introduction to machine learning is beyond the scope of this article, but several are available elsewhere.^{46,47}

Potential Limitations of Machine Learning

We encourage investigators evaluating the impact of machine learning models to keep several considerations in mind. First, machine learning originated in the computer science and engineering communities, but medical machine learning models will most likely operate in either a diagnostic or prognostic setting. Therefore, the recommendations from the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement⁴⁸ are appropriate. If an article does not report results from a new machine learning model in a manner consistent with the TRIPOD statement, readers should interpret the results with caution.

Second, model interpretability is a critical challenge for machine learning in general⁴⁹ and for clinical decision support development in particular.⁵⁰ For clinical decision support tools to be useful, clinicians will need to trust them. Black box models are less likely to be trusted.⁵¹ Visualization of the most important variables supporting a prediction may be useful.⁵² All biomedical machine learning projects require some level of collaboration with domain experts to increase the likelihood of successful implementation of the models.⁵³

Finally, machine learning methods often require analyst input to set model parameters and choose variables and their form (known as feature engineering). These choices, which may not always be transparently reported or reproducible, can influence the performance of the models and the down-stream clinical impact. One example is that unsupervised clustering methods can be prone to bias because selecting the optimal number of clusters typically requires analyst input.

Transfer Learning and Reinforcement Learning

Two more recent innovations within machine learning may have implications for pediatric research (Figure). First, owing to the relative lack of large, high-quality shared data sources in pediatrics, we anticipate that transfer learning⁵⁴ will be very useful to child health investigators. Fewer training data may be required because prediction models from other sources (eg, models trained on adult data) are transferred to pediatric data and then tuned to the new setting.⁵⁵ Second, in most medical settings, the management of pediatric patients is a dynamic process with sequences of decisions by providers and families. Reinforcement learning,⁵⁶ in which computer programs make sequences of actions to maximize a cumulative reward, may be a useful tool as investigators attempt to identify optimal clinical decision sequences.^{57,58}

Deep Learning

Deep learning, a form of machine learning, has shown remarkable accuracy in classification tasks, particularly those with voluminous amounts of biomedical imaging data available for

training (Figure).⁵⁹ Pediatric applications of deep learning include accurate estimations of bone age from hand radiographs^{60,61} and the prediction of the need for infant cerebrospinal fluid diversion from fetal magnetic resonance imaging.⁶² At a high level, deep learning consists of fitting nonlinear models to data with the goal of predicting an outcome using input variables. These neural network models transform the data through a series of hidden layers. The number of hidden layers and connections between the layers in conjunction with the functions used to pool the input data constitute the architecture of the neural network.⁶³

Because of the nonlinear nature and the large number of variables included in deep learning models, there is a high risk of overfitting them to biomedical datasets. Deep learning can be either supervised or unsupervised (Figure), but its greatest successes have been for image recognition problems with gigabytes of data available for training. This is not typical for clinical and biomedical research. To apply deep learning, data with labels indicating the outcome of interest are often required (for supervised learning). Substantial effort by domain experts may be required to label, for example, tumor cells and normal cells in a pathologic image, stroke region boundaries on a radiologic image, or patients with the diagnosis of interest in a corpus of text notes. Despite these challenges, we expect the use of deep learning models to continue to expand in child health research, particularly for pathologic and radiologic images.

Examples of Data Science Tasks in Child Health Research

Phenotyping

Phenotypes are “measurable biological (physiological, biochemical, and anatomical features), behavioral (psychometric pattern), or cognitive markers that are found more often in individuals with a disease or condition than in the general population.”⁶⁴ Computable phenotypes contain computer-readable definitions for identifying patients with certain diseases using known feature variables (eg, specific diagnostic codes, genomic signatures, prescribed medications, or laboratory results within a certain range) characteristics of a specific disease.^{65–67} A number of federal agencies have supported computable phenotype development including the Centers for Medicare and Medicaid Services,^{68,69} the Agency for Healthcare Research and Quality,⁷⁰ the Electronic Medical Records and Genomics Network,⁷¹ the Office of the National Coordinator for Health Information Technology,^{72,73} and the US Food and Drug Administration.⁷⁴ Computable phenotypes are often developed using data from clinical trials, clinical data warehouses, and EHRs because they contain a wealth of patient information (eg, diagnoses, prescribed medications, laboratory measurements, procedures, physical examinations and assessments, and physician narratives).^{75,76}

Computational phenotyping techniques have been applied to many pediatric disorders including encephalopathies,⁷⁷ insomnia,⁷⁸ rare disease,^{79,80} chronic obstructive airway disease,⁸¹ pulmonary vascular disease,⁸² obesity,⁸³ cancer,^{84,85} autism,^{86,87} asthma,^{88,89} acute respiratory distress syndrome (ARDS),⁹⁰ and sepsis.⁹¹ We focus on the last 3 as exemplars.

Phenotypes

Asthma, ARDS, and sepsis are heterogeneous and complex diseases composed of several subphenotypes.^{92–98} Machine learning-based subtyping techniques have been applied to all 3 diseases. Some studies have used unsupervised learning to identify novel clusters or subgroups.^{93,94,99} In ARDS, the most cited work used latent class analysis,¹⁰⁰ an approach that, in a manner similar to unsupervised learning, searches for previously unknown subgroups defined by an unobserved variable.^{90,98,101,102} This approach has also been used with asthma phenotyping.¹⁰³ In contrast with asthma and ARDS, clustering in sepsis has been performed as a feature selection method before classification.⁹⁹

Investigators have then used supervised learning methods including logistic regression^{95,104} and random forest,¹⁰⁵ a tree-based statistical learning method, to test for associations between membership in novel subgroups and outcomes.^{96,97,104}

Limitations of Phenotyping

Commonly applied phenotyping approaches have limitations. First, for both asthma and ARDS, many phenotypes have been developed using clinical trial data. Homogeneity among patients in efficacy clinical trials may be a limitation of those phenotypes.¹⁰⁶ Second, most pediatric phenotyping to date has used only clinical or only molecular data, but not both. Future pediatric phenotyping should integrate multiple heterogeneous sources of patient data (eg, clinical, molecular, environmental, social determinants of health, social media, and other data) to provide a deeper understanding of an individual patient's molecular drivers of disease¹⁰⁷ within the context of their phenotype.

Predictive Analytics Using Frequently Measured Data

Predictive analytic tools use current or historical data to predict future events or behaviors.¹⁰⁸ Applications of predictive analytics in healthcare include risk assessment, clinical decision support, home health monitoring, and resource allocation.¹⁰⁹ Frequently updating predictive analytic tools attempt to use physiologic and biochemical data to identify imminent deterioration or signatures of particular diseases early enough in a patient's course that interventions can improve outcome.^{110,111} Scores or indices derived from these algorithms can be displayed to alert care providers to patients at risk. In adult patients, algorithms have been developed that accurately predict events such as sepsis, acute hemorrhage, urgent unplanned intubation, and urgent unplanned ICU transfer with up to 24 hours of advance warning.^{112–119}

Pediatric predictive analytics using frequently measured variables have largely focused on patients in the neonatal ICU and pediatric ICU. They are more likely to be monitored continuously (often invasively) than other pediatric inpatients, providing the highest resolution of physiologic data. The ability to link continuous bedside monitoring streams with EHRs and other data sources in real time or near real time makes informatics support for time-critical clinical decisions possible.¹²⁰ In a multicenter, randomized trial of more than 3000 very low birth weight infants, displaying a sepsis risk score based on heart rate characteristics was associated with a 21% decrease in mortality, with a number needed to

monitor of 48 infants to prevent 1 mortality.¹²¹ In addition to sepsis, algorithms identifying neonates at risk for necrotizing enterocolitis and apnea have also been developed.^{122,123}

Because of wide ranges of patient age and pathology, both normal and abnormal ICU physiologic measurements have substantial variability. This variability poses challenges for algorithm development. Although no large-scale, publicly available resource such as the MIMIC exists in child health research, several groups have recently reported the use of complex, frequently measured pediatric data streams.^{124–126} Published distributions and patterns for high-frequency vital signs stratified by age¹²⁷ and diagnosis¹²⁸ will be useful for investigators building clinical decision support tools and specifying thresholds for intervention and targeted outcomes.

Predictive analytics using more complex approaches including machine learning are starting to demonstrate promise. In a large cohort of critically ill children, k-means clustering was recently used to develop clusters with separable signatures of illness severity and mortality.¹²⁹ Convolutional neural networks, a form of deep learning, have recently been used to identify physiomarkers that predict severe sepsis in children.¹³⁰ In a cohort of patients with shunted cyanotic heart disease in the cardiac ICU, a classification algorithm built on continuous high-resolution physiologic data can detect impending deterioration events 1–2 hours early.¹³¹

Several limitations are important. All predictive analytic tools will require broad validation studies to demonstrate generalizability and effectiveness. Visualization of complex and dense data streams and presentation of model output to clinicians in ways that facilitate, rather than impede, decision making will also be critical to the success of predictive analytic tools.^{132–134}

Diagnosis and Therapy of Rare and/or Congenital Diseases

One area within pediatrics that has undoubtedly benefitted from advances in data science methods and tools is the genetic diagnosis of children with rare, severe, previously unexplained congenital diseases.^{135–139} Such patients may display symptom constellations ranging from developmental delay and other neuropsychological manifestations to an array of physical birth defects and metabolic derangements. Interestingly, other EHR data (eg, physical examination findings, laboratory values, text assessments) may generate signatures that contribute to patients being suspected of an undiagnosed syndrome.

WES sequences the protein-coding regions that make up approximately 1.5% of the human genome. The first report of a diagnosis made using WES, a mutation in an anion exchange protein leading to congenital chloride diarrhea, was less than a decade ago.¹⁴⁰ Since then, WES has rapidly become a standard test when the diagnosis is uncertain and a congenital disease is suspected.¹⁴¹ Despite the cost associated with WES, several studies have shown that it can avoid even more costly diagnostic odysseys.^{142–145} WES will provide a concrete clinical diagnosis, however, only when an anomaly is found among the roughly 5000 of our estimated 20 000 genes that are already defined as disease causing.¹⁴⁶ Because we lack a sufficient understanding of the remaining 75% of human genes, both clinicians and researchers are frequently faced with the question of how to approach a patient with

symptoms and findings suggesting a genetic etiology, but without a definitive clinical diagnosis.

WES has also been useful in research settings. It has facilitated both large cohort studies and analyses of severe, rare diseases in small families or single patients. One example of this flexibility is the application of WES to a large national cohort of children with CHD.^{147–149} Even large cohorts may miss causative mutations that are extremely rare or that exist only in certain subpopulations.¹⁵⁰ In such situations, WES also allows for the meaningful analysis and identification of potential candidate genes (frequently noted as “variants of uncertain significance” on a clinical exome report) for even a single patient by focusing the data analysis on rare recessive or de novo (not present in either parent but arising spontaneously in the child) variants. Importantly, relying on small patient numbers requires reliable mechanistic data and animal modeling to confirm that variants in candidate genes identified by sequencing are in fact causative for CHD.^{151–153} This combined genetic and functional approach has led to the characterization of a surprising array of genes underlying CHD such as nucleoporins and actin modulators.^{154,155} This approach emphasizes the importance of close collaboration between clinicians, geneticists, data analysts, and bench researchers, as well as collaborative networks such as the Matchmaker Exchange that help to bring these parties together.¹⁵⁶

Despite the wide availability of high-quality DNA sequencing, analysis of the data remains a challenge for both clinicians and researchers. Common analysis tools such as Genome Analysis Toolkit require a degree of comfort with the Linux operating system command line and general purpose programming.¹⁵⁷ Work is ongoing, however, on web-based and user-friendly interfaces to make raw WES data analysis possible even for relative computing novices.

Advances in next-generation technologies such as WES and whole genome sequencing have allowed us to move beyond descriptive diagnoses such as tetralogy of Fallot, inflammatory bowel disease, or common variable immune deficiency and begin to provide precise molecular diagnoses.¹⁴⁶ As sequencing costs decrease, whole genome sequencing will begin to replace WES as the tool of choice for genetic analysis of patients.¹⁵⁸ In addition, complementary approaches to NGS such as RNA expression analysis and metabolomics will expand our research capacity.

Pharmacoepidemiology and Pharmacovigilance

Medication data present a unique opportunity for data scientists to leverage multiple clinical data sources to advance pharmacovigilance (monitoring drug effects after licensing). For patient types in whom trials are difficult to conduct (eg, medically complex and critically ill children), real-world post-marketing studies are central to understanding the efficacy and safety of medications.¹⁵⁹ However, most administrative pharmacy databases do not contain all of the necessary components to evaluate the relationship between a medication exposure and an outcome of interest. For example, a potential adverse drug event requires identification of a patient, a possible event, use of a suspected drug, and evidence of an adverse event/outcome.¹⁶⁰ Novel linkages between traditional administrative data and the

various, rich clinical data available through EHRs will improve the ability to answer important medication-related clinical and research questions.

First, medication exposure data can be enhanced through assembly of complete and valid patient medication profiles. Medication profiles may include exposure to single medications, drug-drug combinations, sequences of medications, or even counts of concurrent medications. Each of these types of exposures depends on knowing all medications a patient takes across all settings of care, regardless of insurance type. This has been enabled by the availability of all-payer claims databases, although these resources have proven somewhat cumbersome to access.^{161,162} Even when complete data are available, pharmacy claims data lack the more detailed clinical prescribing information captured in the EHR (eg, linkage to an indicated diagnosis for use of a medication or the medical decision making supporting use of a medication) needed to assess appropriateness of prescribing practices. Many medications are still used off-label in pediatrics (eg, clonidine) and without this clinical information, it is difficult to determine a medication's appropriateness. The successful integration of administrative and clinical data sources will improve the understanding of why a medication was used, which will ultimately benefit bedside care and pharmacovigilance efforts.¹⁶³ Additional patient-level adherence data, such as patient-reported adherence or technology solutions (eg, wi-fi enabled pill bottles) may provide even more valid exposure information.

Next, robust outcome data—which are traditionally lacking in administrative claims aside from what can be gleaned through ICD or Current Procedural Terminology codes—exist in the EHR and should be leveraged. The extraction of text from clinical notes and the subsequent mapping of terms, such as adverse clinical symptoms, to SNOMED-CT would increase the ability to detect therapeutic responses or adverse effects attributable to medications.¹⁶⁴ Limited laboratory and radiographic data are populated for some commercial insurance pharmacy databases (eg, Premier Healthcare Database, Truven MarketScan Database),^{165,166} but not for public insurance databases (eg, Medicaid MAX Databases), which tend to contain more children with medical complexity and increased use of medications.¹⁶⁷ Other novel outcome data with potential applicability to medication-related questions such as patient-reported outcomes^{168,169} or consumer device data (eg, sleep monitoring from wearables) are increasingly available.

Finally, combined clinical, medication exposure, and outcome data will facilitate analyses ranging from traditional pharmacoepidemiology study designs (eg, cohort and case control studies) to newer types of machine learning analyses. For example, to supplement the US Food and Drug Administration's Adverse Event Reporting System, in which the denominator is unknown, algorithmic analysis of large volumes of complete medication data may improve the identification of signals of interest. Data mining techniques, such as the use of association rules, may increase the ability to isolate new meaningful signals from background noise and subsequently prioritize certain medications for further study.¹⁷⁰ Although an increased volume of data will enable the identification of rare events, it will also require advanced approaches for handling confounding in observational studies, methods for longitudinal data analysis of big data, and procedures for exploratory data mining, machine learning, and data visualization.^{171,172}

Conclusions

Data science is beginning to make an impact in child health research and its reach is likely to grow. Pediatric investigators implementing data science techniques have some unique challenges (eg, imperative for data sharing, importance of potentially sensitive temporal information) and risks (eg, compromising data privacy, off-target effects of computational tools¹⁷³). However, they also have opportunities. The relative rarity of chronic medical conditions in childhood may give child health investigators using large collections of existing data an advantage relative to those studying adults. Fewer children than adults have multiple medical conditions, which may make phenotyping and causal inference more achievable. Also, pediatric investigators might lead in the application of data science methods to lifespan diseases such as asthma¹⁷⁴ and epilepsy. Finally, pediatric research has fewer clinical trials than adult research.¹⁷⁵ Data science has the potential to both generate interventions to be trialed in children in the future and also provide actionable evidence on topics with less high-quality evidence to date.

This review was necessarily limited. Comprehensive coverage of any one of the individual fields (eg, machine learning, informatics, statistics, genomics) that contribute to pediatric data science is beyond the scope of this work. We also did not focus on workforce development or data privacy, both of which are timely and critically important topics. Finally, because of its multidisciplinary nature, data science requires study design, statistical, computational, scientific, and/or clinical domain knowledge, and other expertise. Because of this, effective communication and many other team science skills are necessary.

Acknowledgments

Supported, in part, by the National Institutes of Health (NICHD R03 HD094912 [to T.B.] and NCATS UL1 TR002535 [to Ronald Sokol] and NICHD K23 HD091295 [to J.F.]). The authors declare no conflicts of interest.

Glossary

ARDS	Acute respiratory distress syndrome
CDM	Common data model
CHD	Congenital heart disease
EHR	Electronic health record
FAIR	Findable, Accessible, Interoperable, and Re-usable
i2b2	Informatics for Integrating Biology & the Bedside
ICD	<i>International Classification of Diseases</i>
ICU	Intensive care unit
MIMIC	Medical Information Mart for Intensive Care
PEDSNet	National Pediatric Learning Health System

SNOMED-CT	Systematized Nomenclature of Medicine Clinical Terms
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
WES	Whole exome sequence

References

1. Adomavicius G, Tuzhilin A. Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans on Data and Knowledge Engineering* 2005;17: 734–49.
2. Altman MC, Busse WW A Deep dive into asthma transcriptomics. Lessons from U-BIOPRED. *Am J Respir Crit Care Med* 2017;195:1279–80. [PubMed: 28504594]
3. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. 25th International Conference on Neural Information Processing Systems; Dec 12–17, 2011; Granada, Spain.
4. Sanchez-Pinto LN, Luo Y, Churpek MM. Big data and data science in critical care. *Chest* 2018;15:1239–48.
5. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA* 2013;309:1351–2. [PubMed: 23549579]
6. Naylor CD. On the prospects for a (deep) learning health care system. *JAMA* 2018;320:1099–100. [PubMed: 30178068]
7. Wong HR. Intensive care medicine in 2050: precision medicine. *Intensive Care Med* 2017;43:1507–9. [PubMed: 28236258]
8. Zhu L, Zheng JW. Informatics, data science, and artificial intelligence. *JAMA* 2018;320:1103–4. [PubMed: 30326503]
9. Dolley S Big data's role in precision public health. *Front Public Health* 2018;6:68. [PubMed: 29594091]
10. NIH Strategic Plan for Data Science. https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf Accessed October 12, 2018
11. Bennett TD, Spaeder MC, Matos RI, Watson RS, Typpo KV, Khemani RG, et al. Existing data analysis in pediatric critical care research. *Front Pediatr* 2014;2:79. [PubMed: 25121079]
12. Harper SA, Bradley JS, Englund JA, File TM, Gravenstein S, Hayden FG, et al. Seasonal influenza in adults and children—diagnosis, treatment, chemoprophylaxis, and institutional outbreak management: clinical practice guidelines of the Infectious Diseases Society of America. *Clin Infect Dis* 2009;48:1003–32. [PubMed: 19281331]
13. Singer E The measured life. *MIT Technology Review*; 2011.
14. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. [PubMed: 26978244]
15. Bill and Melinda Gates Foundation. Bill and Melinda Gates Foundation Open Access Policy. <https://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy> Accessed October 9, 2018.
16. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035. [PubMed: 27219127]
17. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, et al. PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc* 2014;21:602–6. [PubMed: 24821737]
18. Pediatric Emergency Care Applied Research Network. The Pediatric Emergency Care Applied Research Network (PECARN): rationale, development, and first steps. *Pediatr Emerg Care* 2003;19:185–93. [PubMed: 12813308]

19. Willson DF, Dean JM, Newth C, Pollack M, Anand KJ, Meert K, et al. Collaborative Pediatric Critical Care Research Network (CPCCRN). *Pediatr Crit Care Med* 2006;7:301–7. [PubMed: 16738501]
20. Srivastava R, Landrigan CP. Development of the Pediatric Research in Inpatient Settings (PRIS) Network: lessons learned. *J Hosp Med* 2012;7: 661–4. [PubMed: 23033225]
21. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010;17: 124–30. [PubMed: 20190053]
22. Weber SC, Lowe H, Das A, Ferris T. A simple heuristic for blindfolded record linkage. *J Am Med Inform Assoc* 2012;19:e157–61. [PubMed: 22298567]
23. Aplenc R, Fisher BT, Huang YS, Li Y, Alonzo TA, Gerbing RB, et al. Merging of the National Cancer Institute-funded cooperative oncology group data with an administrative data source to develop a more effective platform for clinical trial analysis and comparative effectiveness research: a report from the Children's Oncology Group. *Pharmacoepidemiol Drug Saf* 2012;21(Suppl 2):37–43. [PubMed: 22552978]
24. Pasquali SK, Jacobs JP, Shook GJ, O'Brien SM, Hall M, Jacobs ML, et al. Linking clinical registry data with administrative data using indirect identifiers: implementation and validation in the congenital heart surgery population. *Am Heart J* 2010;160:1099–104. [PubMed: 21146664]
25. Bennett TD, Dean JM, Keenan HT, McGlinchey MH, Thomas AM, Cook LJ. Linked records of children with traumatic brain injury. Probabilistic linkage without use of protected health information. *Methods InfMed* 2015;54:328–37.
26. Belle A, Thiagarajan R, Soroushmehr SM, Navidi F, Beard DA, Najarian K. Big data analytics in healthcare. *Biomed Res Int* 2015;2015:370194. [PubMed: 26229957]
27. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014;2:3. [PubMed: 25825667]
28. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *J Am Med Inform Assoc* 2009;16:316–27. [PubMed: 19261934]
29. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc* 2007;14:687–96. [PubMed: 17712081]
30. Schleif F, Brusselle G, Louis R, Vandenplas O, Michils A, Pilette C, et al. Heterogeneity of phenotypes in severe asthmatics. The Belgian Severe Asthma Registry (BSAR). *Respir Med* 2014;108:1723–32. [PubMed: 25456708]
31. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. *Med Care* 2012;50(Suppl):S60–7. [PubMed: 22692260]
32. Humphreys BL, Lindberg DA. The UMLS project: making the conceptual connection between users and the information they need. *Bull Med Libr Assoc* 1993;81:170–7. [PubMed: 8472002]
33. Wenzel SE. Characteristics, definition and phenotypes of severe asthma. In: Chung KF, Bel EH, Wenzel SE, eds. *Difficult-to-Treat Severe Asthma*. European Respiratory Society Monographs; 2011 <https://books.ersjournals.com/content/difficult-to-treat-severe-asthma.tab-info> Accessed January 15, 2019.
34. Kahn MG, Bailey LC, Forrest CB, Padula MA, Hirschfeld S. Building a common pediatric research terminology for accelerating child health research. *Pediatrics* 2014;133:516–25. [PubMed: 24534404]
35. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. *AMIA Jt Summits Transl Sci Proc* 2010;2010:1–5.
36. Capurro D, Yetisgen M, van Eaton E, Black R, Tarczy-Hornoch P. Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multisite assessment. *EGEMS (Wash DC)* 2014;2:1079. [PubMed: 25848594]
37. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc* 2002;9:600–11. [PubMed: 12386111]
38. Office of the National Coordinator (ONC). Interoperability Roadmap. www.HealthIT.gov Accessed May 28, 2018.

39. Ong T, Pradhananga R, Holve E, Kahn MG. A framework for classification of electronic health data extraction-transformation-loading challenges in data network participation. EGEMS (Wash DC) 2017;5:10. [PubMed: 29930958]
40. Huser V, Cimino JJ. Desiderata for healthcare integrated data repositories based on architectural comparison of three public repositories. AMIA Annu Symp Proc 2013;2013:648–56. [PubMed: 24551366]
41. Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G, et al. The HMO Research network virtual data warehouse: a public data model to support collaboration. EGEMS (Wash DC) 2014;2:1049. [PubMed: 25848584]
42. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19:54–60. [PubMed: 22037893]
43. Qualls LG, Phillips TA, Hammill BG, Topping J, Louzao DM, Brown JS, et al. Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet(R)). EGEMS (Wash DC) 2018;6:3. [PubMed: 29881761]
44. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484–9. [PubMed: 26819042]
45. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature 2017;550:354–9. [PubMed: 29052630]
46. Deo RC. Machine learning in medicine. Circulation 2015;132:1920–30. [PubMed: 26572668]
47. Bishop C. Pattern recognition and machine learning. Berlin: Springer-Verlag; 2006.
48. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1–73. [PubMed: 25560730]
49. Doshi-Velez F, Kim B. Towards a rigorous science on interpretable machine learning. <https://arxiv.org/abs/1702.08608> Accessed April 29, 2018.
50. Matheny ME. Generation of Knowledge for Clinical Decision Support In: Greenes RA ed. Clinical decision support: the road to broad adoption. 2nd ed. Philadelphia: Elsevier; 2014 p. 309–37.
51. Shortliffe EH, Sepulveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018;320:2199–200. [PubMed: 30398550]
52. Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?”: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining; August 13–17, 2016; San Francisco, CA.
53. Meyer NJ, Calfee CS. Novel translational approaches to the search for precision therapies for acute respiratory distress syndrome. Lancet Respir Med 2017;5:512–23. [PubMed: 28664850]
54. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22:1345–59.
55. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172:1122–31.e9. [PubMed: 29474911]
56. Sutton R, Barto A. Reinforcement learning: an introduction. Cambridge, MA: MIT Press/Bradford Books; 1998.
57. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518:529–33. [PubMed: 25719670]
58. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018;24:1716–20. [PubMed: 30349085]
59. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10. [PubMed: 27898976]
60. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018;287:313–22. [PubMed: 29095675]

61. Mutasa S, Chang PD, Ruzal-Shapiro C, Ayyala R. MABAL: a novel deep-learning architecture for machine-assisted bone age labeling. *J Digit Imaging* 2018;31:513–9. [PubMed: 29404850]
62. Pisapia JM, Akbari H, Rozycki M, Goldstein H, Bakas S, Rathore S, et al. Use of fetal magnetic resonance image analysis and machine learning to predict the need for postnatal cerebrospinal fluid diversion in fetal ventriculomegaly. *JAMA Pediatr* 2018;172:128–35. [PubMed: 29255892]
63. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44. [PubMed: 26017442]
64. Electronic health records-based phenotyping. Rethinking clinical trials[®]. <https://sites.duke.edu/rethinkingclinicaltrials/ehr-phenotyping/> Accessed August 10, 2018.
65. Lussier YA, Liu Y. Computational approaches to phenotyping: high-throughput phenomics. *Proc Am Thorac Soc* 2007;4:18–25. [PubMed: 17202287]
66. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014;21:221–30. [PubMed: 24201027]
67. Bennett TD, DeWitt PE, Dixon RR, Kartchner C, Sierra Y, Ladell D, et al. Development and prospective validation of tools to accurately identify neurosurgical and critical care events in children with traumatic brain injury. *Pediatr Crit Care Med* 2017;18:442–51. [PubMed: 28252524]
68. Centers for Medicare and Medicaid Services. QualityNet, <https://www.qualitynet.org/> Accessed August 10, 2018.
69. Centers For Medicare and Medicaid Services. Chronic conditions data warehouse: condition categories. <https://www.ccwdata.org/web/guest/condition-categories> Accessed October 12, 2018.
70. HCUP-US Tools & Software Page. <http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp> Accessed August 10, 2018.
71. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc* 2016;23: 1046–52. [PubMed: 27026615]
72. SHARP (Strategic Health IT Advanced Research Projects) Program, [HealthIT.gov. http://www.healthit.gov/policy-researchers-implementers/strategic-health-it-advanced-research-projects-sharp](http://www.healthit.gov/policy-researchers-implementers/strategic-health-it-advanced-research-projects-sharp) Accessed August 10, 2018.
73. Bodenreider O, Nguyen D, Chiang P, Chuang P, Madden M, Winnenburg R, et al. The NLM value set authority center. *Stud Health Technol Inform* 2013;192:1224. [PubMed: 23920998]
74. U.S. Food and Drug Administration. FDA's sentinel initiative. <https://www.sentinelinitiative.org/> Accessed October 12, 2018.
75. Denny JC. Mining electronic health records in the genomics era. *PLoS Comput Biol* 2012;8:e1002823. [PubMed: 23300414]
76. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med* 2015;7:41. [PubMed: 25937834]
77. Ma M, Adams HR, Seltzer LE, Dobyns WB, Paciorkowski AR. Phenotype differentiation of FOXG1 and MECP2 disorders: a new method for characterization of developmental encephalopathies. *J Pediatr* 2016;178:233–40.e10. [PubMed: 27640358]
78. Bruni O, Sette S, Angriman M, Baumgartner E, Selvaggini L, Belli C, et al. Clinically oriented subtyping of chronic insomnia of childhood. *J Pediatr* 2018;196:194–200.e1. [PubMed: 29550236]
79. Michalik DE, Taylor BW, Panepinto JA. Identification and validation of a sickle cell disease cohort within electronic health records. *Acad Pediatr* 2017;17:283–7. [PubMed: 27979750]
80. Son JH, Xie G, Yuan C, Ena L, Li Z, Goldstein A, et al. Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. *Am J Hum Genet* 2018;103:58–73. [PubMed: 29961570]
81. Nyilas S, Singer F, Kumar N, Yammine S, Meier-Girard D, Koerner-Rettberg C, et al. Physiological phenotyping of pediatric chronic obstructive airway diseases. *J Appl Physiol* 2016;121:324–32. [PubMed: 27231309]
82. Goss KN, Everett AD, Mourani PM, Baker CD, Abman SH. Addressing the challenges of phenotyping pediatric pulmonary vascular disease. *Pulm Circ* 2017;7:7–19. [PubMed: 28680562]

83. Lingren T, Thaker V, Brady C, Namjou B, Kennebeck S, Bickel J, et al. Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. *Appl Clin Inform* 2016;7:693–706. [PubMed: 27452794]
84. Tung WL, Quek C. GenSo-FDSS: a neural-fuzzy decision support system for pediatric ALL cancer subtype identification using gene expression data. *Artif Intell Med* 2005;33:61–88. [PubMed: 15617982]
85. Obulkasim A, Fornerod M, Zwaan MC, Reinhardt D, van den Heuvel-Eibrink MM. Subtype prediction in pediatric acute myeloid leukemia: classification using differential network rank conservation revisited. *BMC Bioinformatics* 2015;16:305. [PubMed: 26399969]
86. Arnett AB, Trinh S, Bernier RA. The state of research on the genetics of autism spectrum disorder: methodological, clinical and conceptual progress. *Curr Opin Psychol* 2018;27:1–5. [PubMed: 30059871]
87. Zwaigenbaum L, Penner M. Autism spectrum disorder: advances in diagnosis and evaluation. *BMJ* 2018;361:k1674. [PubMed: 29784657]
88. Ross MK, Yoon J, van der Schaar A, van der Schaar M. Discovering pediatric asthma phenotypes on the basis of response to controller medication using machine learning. *Ann Am Thorac Soc* 2018;15:49–58. [PubMed: 29048949]
89. Chung KF. Asthma phenotyping: a necessity for improved therapeutic precision and new targeted therapies. *J Intern Med* 2016;279:192–204. [PubMed: 26076339]
90. Calfee CS, Delucchi K, Parsons PE, Thompson BT, Ware LB, Matthay MA, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir Med* 2014;2:611–20. [PubMed: 24853585]
91. Mickiewicz B, Thompson GC, Blackwood J, Jenne CN, Winston BW, Vogel HJ, et al. Development of metabolic and inflammatory mediator biomarker phenotyping for early diagnosis and triage of pediatric sepsis. *Crit Care* 2015;19:320. [PubMed: 26349677]
92. Moore WC, Bleecker ER, Curran-Everett D, Erzurum SC, Ameredes BT, Bacharier L, et al. Characterization of the severe asthma phenotype by the National Heart, Lung, and Blood Institute's Severe Asthma Research Program. *J Allergy Clin Immunol* 2007;119:405–13. [PubMed: 17291857]
93. Bos LD, Schouten LR, van Vught LA, Wiewel MA, Ong DSY, Cremer O, et al. Identification and validation of distinct biological phenotypes in patients with acute respiratory distress syndrome by cluster analysis. *Thorax* 2017;72:876–83. [PubMed: 28450529]
94. Wong HR, Wheeler DS, Tegtmeier K, Poynter SE, Kaplan JM, Chima RS, et al. Toward a clinically feasible gene expression-based sub-classification strategy for septic shock: proof of concept. *Crit Care Med* 2010;38:1955–61. [PubMed: 20639748]
95. Wong HR, Cvijanovich NZ, Anas N, Allen GL, Thomas NJ, Bigham MT, et al. Developing a clinically feasible personalized medicine approach to pediatric septic shock. *Am J Respir Crit Care Med* 2015;191:309–15. [PubMed: 25489881]
96. Brown LM, Calfee CS, Matthay MA, Brower RG, Thompson BT, Checkley W, et al. A simple classification model for hospital mortality in patients with acute lung injury managed with lung protective ventilation. *Crit Care Med* 2011;39:2645–51. [PubMed: 21725235]
97. Just J, Gouvis-Echraghi R, Couderc R, Guillemot-Lambert N, Saint-Pierre P. Novel severe wheezy young children phenotypes: boys atopic multiple-trigger and girls nonatopic uncontrolled wheeze. *J Allergy Clin Immunol* 2012;130:103–10.e8. [PubMed: 22502798]
98. Shankar-Hari M, McAuley DF. Acute respiratory distress syndrome phenotypes and identifying treatable traits. the dawn of personalized medicine for ARDS. *Am J Respir Crit Care Med* 2017;195:280–1. [PubMed: 28145757]
99. Wong HR, Cvijanovich N, Lin R, Allen GL, Thomas NJ, Willson DF, et al. Identification of pediatric septic shock subclasses based on genome-wide expression profiling. *BMC Med* 2009;7:34. [PubMed: 19624809]
100. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974;61:215–31.
101. Calfee CS, Delucchi KL, Sinha P, Matthay MA, Hackett J, Shankar-Hari M, et al. Acute respiratory distress syndrome subphenotypes and differential response to simvastatin: secondary

- analysis of a randomised controlled trial. *Lancet Respir Med* 2018;6:691–8. [PubMed: 30078618]
102. Zhang Z Identification of three classes of acute respiratory distress syndrome using latent class analysis. *Peer J* 2018;6:e4592. [PubMed: 29610712]
 103. Deliu M, Sperrin M, Belgrave D, Custovic A. Identification of asthma subtypes using clustering methodologies. *Pulm Ther* 2016;2:19–41. [PubMed: 27512723]
 104. Yehya N, Keim G, Thomas NJ. Subtypes of pediatric acute respiratory distress syndrome have different predictors of mortality. *Intensive Care Med* 2018;44:1230–9. [PubMed: 29971591]
 105. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, NY: Springer; 2009.
 106. Belgrave D, Henderson J, Simpson A, Buchan I, Bishop C, Custovic A. Disaggregating asthma: big investigation versus big data. *J Allergy Clin Immunol* 2017;139:400–7. [PubMed: 27871876]
 107. Dudley JT, Listgarten J, Stegle O, Brenner SE, Parts L. Personalized medicine: from genotypes, molecular phenotypes and the quantified self, towards improved medicine. *Pac Symp Biocomput* 2015:342–6.
 108. Nyce C, American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America. Predictive analytics white paper. <https://www.the-digital-insurer.com/wp-content/uploads/2013/12/78-Predictive-Modeling-White-Paper.pdf> Accessed January 15, 2019.
 109. Suresh S Big data and predictive analytics: applications in the care of children. *Pediatr Clin North Am* 2016;63:357–66. [PubMed: 27017041]
 110. Moorman JR, Rusin CE, Lee H, Guin LE, Clark MT, Delos JB, et al. Predictive monitoring for early detection of subacute potentially catastrophic illnesses in critical care. *Conf Proc IEEE Eng Med Biol Soc* 2011;2011:5515–8. [PubMed: 22255587]
 111. Fairchild KD. Predictive monitoring for early detection of sepsis in neonatal ICU patients. *Curr Opin Pediatr* 2013;25:172–9. [PubMed: 23407184]
 112. Moss TJ, Lake DE, Calland JF, Enfield KB, Delos JB, Fairchild KD, et al. Signatures of subacute potentially catastrophic illness in the ICU: model development and validation. *Crit Care Med* 2016;44: 1639–48. [PubMed: 27452809]
 113. Moss TJ, Clark MT, Calland JF, Enfield KB, Voss JD, Lake DE, et al. Cardiorespiratory dynamics measured from continuous ECG monitoring improves detection of deterioration in acute care patients: a retrospective cohort study. *PLoS One* 2017;12:e0181448. [PubMed: 28771487]
 114. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open* 2018;8:e017833.
 115. Politano AD, Riccio LM, Lake DE, Rusin CG, Guin LE, Josef CS, et al. Predicting the need for urgent intubation in a surgical/trauma intensive care unit. *Surgery* 2013;154:1110–6. [PubMed: 24075272]
 116. Moss TJ, Clark MT, Lake DE, Moorman JR, Calland JF. Heart rate dynamics preceding hemorrhage in the intensive care unit. *J Electrocardiol* 2015;48:1075–80. [PubMed: 26342251]
 117. Hooper MH, Weavind L, Wheeler AP, Martin JB, Gowda SS, Semler MW, et al. Randomized trial of automated, electronic monitoring to facilitate early detection of sepsis in the intensive care unit*. *Crit Care Med* 2012;40:2096–101. [PubMed: 22584763]
 118. Sawyer AM, Deal EN, Labelle AJ, Witt C, Thiel SW, Heard K, et al. Implementation of a real-time computerized sepsis alert in nonintensive care unit patients. *Crit Care Med* 2011;39:469–73. [PubMed: 21169824]
 119. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016;23:269–78. [PubMed: 26679719]
 120. De Georgia MA, Kaffashi F, Jacono FJ, Loparo KA. Information technology in critical care: review of monitoring and data acquisition systems for patient care and research. *Sci World J* 2015;2015: 727694.
 121. Moorman JR, Carlo WA, Kattwinkel J, Schelonka RL, Porcelli PJ, Navarrete CT, et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial. *J Pediatr* 2011;159:900–6.e1. [PubMed: 21864846]

122. Fairchild KD, Lake DE. Cross-correlation of heart rate and oxygen saturation in very low birthweight infants: association with apnea and adverse events. *Am J Perinatol* 2018;35:463–9. [PubMed: 29141263]
123. Fairchild KD, Lake DE, Kattwinkel J, Moorman JR, Bateman DA, Grieve PG, et al. Vital signs and their cross-correlation in sepsis and NEC: a study of 1,065 very-low-birth-weight infants in two NICUs. *Pediatr Res* 2017;81:315–21. [PubMed: 28001143]
124. Blount M, McGregor C, James A, Sow D, Kamaleswaran R, Tuuha S, et al. On the integration of an artifact system and a real-time healthcare analytics system. *Proceedings of the 1st ACM International Health Informatics Symposium*; November 11–12, 2010 Arlington, VA p. 647–55.
125. Brossier D, El Taani R, Sauthier M, Roumeliotis N, Emeriaud G, Jouvet P. Creating a high-frequency electronic database in the PICU: the perpetual patient. *Pediatr Crit Care Med* 2018;19:e189–98. [PubMed: 29406373]
126. Wetzel RC. Pediatric intensive care databases for quality improvement. *J Pediatr Intensive Care* 2016;5:81–8.
127. Eytan D, Goodwin AJ, Greer R, Guerguerian AM, Laussen PC. Heart rate and blood pressure centile curves and distributions by age of hospitalized critically ill children. *Front Pediatr* 2017;5:52. [PubMed: 28367430]
128. Eytan D, Goodwin AJ, Greer R, Guerguerian AM, Mazwi M, Laussen PC. Distributions and behavior of vital signs in critically ill children by admission diagnosis. *Pediatr Crit Care Med* 2018;19: 115–24. [PubMed: 29206728]
129. Williams JB, Ghosh D, Wetzel RC. Applying machine learning to pediatric critical care data. *Pediatr Crit Care Med* 2018;19:599–608. [PubMed: 29727354]
130. Kamaleswaran R, Akbilgic O, Hallman MA, West AN, Davis RL, Shah SH. Applying artificial intelligence to identify physiologic markers predicting severe sepsis in the PICU. *Pediatr Crit Care Med* 2018;19:e495–503. [PubMed: 30052552]
131. Rusin CG, Acosta SI, Shekerdemian LS, Vu EL, Bavare AC, Myers RB, et al. Prediction of imminent, severe deterioration of children with parallel circulations using real-time processing of physiologic data. *J Thorac Cardiovasc Surg* 2016;152:171–7. [PubMed: 27174513]
132. Segall N, Borbolla D, Del Fiore G, Waller R, Reese T, Nesbitt P, et al. Trend displays to support critical care: a systematic review. *IEEE International Conference on Healthcare Informatics (ICHI)*; August 23–26, 2017 Park City, UT p. 305–13.
133. West VL, Borland D, Hammond WE. Innovative information visualization of electronic health record data: a systematic review. *J Am Med Inform Assoc* 2015;22:330–9. [PubMed: 25336597]
134. Gorges M, Staggers N. Evaluations of physiological monitoring displays: a systematic review. *J Clin Monit Comput* 2008;22:45–66. [PubMed: 18064532]
135. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;12:745–55. [PubMed: 21946919]
136. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* 2011;3:65ra4.
137. Ng SB, Nickerson DA, Bamshad MJ, Shendure J. Massively parallel sequencing and rare disease. *Hum Mol Genet* 2010;19:R119–24. [PubMed: 20846941]
138. Rabbani B, Mahdih N, Hosomichi K, Nakaoka H, Inoue I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *J Hum Genet* 2012;57:621–32. [PubMed: 22832387]
139. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* 2013;369:1502–11. [PubMed: 24088041]
140. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 2009;106:19096–101. [PubMed: 19861545]

141. Smith HS, Swint JM, Lalani SR, Yamal JM, Otto MC de Oliveira, Castellanos S, et al. Clinical application of genome and exome sequencing as a diagnostic tool for pediatric patients: a scoping review of the literature. *Genet Med* 2019;21:3–16. [PubMed: 29760485]
142. Cordoba M, Rodriguez-Quiroga SA, Vega PA, Salinas V, Perez-Maturo J, Amartino H, et al. Whole exome sequencing in neurogenetic odysseys: an effective, cost- and time-saving diagnostic approach. *PLoS One* 2018;13:e0191228. [PubMed: 29389947]
143. Monroe GR, Frederix GW, Savelberg SM, de Vries TI, Duran KJ, van der Smagt JJ, et al. Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. *Genet Med* 2016;18:949–56. [PubMed: 26845106]
144. Vrijenhoek T, Middelburg EM, Monroe GR, van Gassen KLI, Geenen JW, Hovels AM, et al. Whole-exome sequencing in intellectual disability; cost before and after a diagnosis. *Eur J Hum Genet* 2018;26:1566–71. [PubMed: 29959382]
145. Tan TY, Dillon OJ, Stark Z, Schofield D, Alam K, Shrestha R, et al. Diagnostic impact and cost-effectiveness of whole-exome sequencing for ambulant children with suspected monogenic conditions. *JAMA Pediatr* 2017;171:855–62. [PubMed: 28759686]
146. Caspar SM, Dubacher N, Kopps AM, Meienberg J, Henggeler C, Matyas G. Clinical sequencing: from raw data to diagnosis with lifetime value. *Clin Genet* 2018;93:508–19. [PubMed: 29206278]
147. Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat Genet* 2017;49:1593–601. [PubMed: 28991257]
148. Homsy J, Zaidi S, Shen Y, Ware JS, Samocha KE, Karczewski KJ, et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science (New York, NY)* 2015;350:1262–6.
149. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, et al. De novo mutations in histone-modifying genes in congenital heart disease. *Nature* 2013;498:220–3. [PubMed: 23665959]
150. Yuan S, Zaidi S, Brueckner M. Congenital heart disease: emerging themes linking genetics and development. *Curr Opin Genet Dev* 2013;23:352–9. [PubMed: 23790954]
151. Liu X, Kim AJ, Reynolds W, Wu Y, Lo CW. Phenotyping cardiac and structural birth defects in fetal and newborn mice. *Birth Defects Res* 2017;109:778–90. [PubMed: 28544620]
152. Grant MG, Patterson VL, Grimes DT, Burdine RD. Modeling syndromic congenital heart defects in zebrafish. *Curr Top Dev Biol* 2017;124:1–40. [PubMed: 28335857]
153. Duncan AR, Khokha MK. *Xenopus* as a model organism for birth defects-congenital heart disease and heterotaxy. *Semin Cell Dev Biol* 2016;51:73–9. [PubMed: 26910255]
154. Kulkarni SS, Griffin JN, Date PP, Liem KF Jr, Khokha MK. WDR5 Stabilizes Actin Architecture to promote multiciliated cell formation. *Dev Cell* 2018;46:595–610.e3. [PubMed: 30205038]
155. Del Viso F, Huang F, Myers J, Chalfant M, Zhang Y, Reza N, et al. Congenital heart disease genetics uncovers context-dependent organization and function of nucleoporins at cilia. *Dev Cell* 2016;38:478–92. [PubMed: 27593162]
156. Sobreira NLM, Arachchi H, Buske OJ, Chong JX, Hutton B, Foreman J, et al. Matchmaker exchange. *Curr Protoc Hum Genet* 2017;95:9.31.1–9.31.15.
157. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303. [PubMed: 20644199]
158. Service RF. Gene sequencing. The race for the \$1000 genome. *Science (New York, NY)* 2006;311:1544–6.
159. McMahon AW, Dal Pan G. Assessing drug safety in children - the role of real-world data. *N Engl J Med* 2018;378:2155–7. [PubMed: 29874532]
160. US Food and Drug Administration. Individual Case Safety Reports. <https://www.fda.gov/forindustry/datastandards/individualcasesafetyreports/default.htm> Accessed October 12, 2018.
161. Agency for Healthcare Research and Quality. Data Evaluation of All-Payer Claims Databases. <http://www.ahrq.gov/professionals/quality-patient-safety/quality-resources/apcd/backgroundppt/data.html> Accessed October 12, 2018.

162. Doshi JA, Hendrick FB, Graff JS, Stuart BC. Data, data everywhere, but access remains a big issue for researchers: a review of access policies for publicly-funded patient-level health care data in the United States. *EGEMS (Wash DC)* 2016;4:1204. [PubMed: 27141517]
163. Comer D, Couto J, Aguiar R, Wu P, Elliott D. Using aggregated pharmacy claims to identify primary nonadherence. *Am J Manag Care* 2015;21:e655–60. [PubMed: 26760428]
164. Wang L, Rastegar-Mojarad M, Liu S, Zhang H, Liu H. Discovering adverse drug events combining spontaneous reports with electronic medical records: a case study of conventional DMARDs and biologics for rheumatoid arthritis. *AMIA Jt Summits Transl Sci Proc* 2017;2017:95–103. [PubMed: 28815115]
165. Premier Inc. White Paper: Premier Hospital Database (PHD) - 2 18, 2018 <https://learn.premierinc.com/i/790965-premier-healthcare-database-whitepaper/1?> Accessed October 12, 2018.
166. Hansen L, Health T. The Truven Health MarketScan Databases for life sciences researchers, 3 2017 <https://truvenhealth.com/Portals/0/Assets/2017-MarketScan-Databases-Life-Sciences-Researchers-WP.pdf> Accessed October 12, 2018.
167. Cohen E, Hall M, Lopert R, Bruen B, Chamberlain LJ, Bardach N, et al. High-expenditure pharmaceutical use among children in Medicaid. *Pediatrics* 2017;140.
168. HealthMeasures. Available PROMIS® Measures for Pediatric Self-Report (ages 8-17) and Parent Proxy Report (ages 5-17). <http://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis/list-of-pediatric-measures> Accessed October 12, 2018
169. Wolfe J, Orellana L, Ullrich C, Cook EF, Kang TI, Rosenberg A, et al. Symptoms and distress in children with advanced cancer: prospective patient-reported outcomes from the PediQUEST Study. *J Clin Oncol* 2015;33:1928–35. [PubMed: 25918277]
170. Feinstein JA, Morrato EH, Feudtner C. Prioritizing pediatric drug research using population-level health data. *JAMA Pediatr* 2017;171:7–8. [PubMed: 27893067]
171. Bate A, Reynolds RF, Cabel P. The hope, hype and reality of big data for pharmacovigilance. *Ther Adv Drug Saf* 2018;9:5–11. [PubMed: 29318002]
172. Bate A Guidance to reinforce the credibility of health care database studies and ensure their appropriate impact. *Pharmacoepidemiol Drug Saf* 2017;26:1013–7. [PubMed: 28913965]
173. O'Neil C Weapons of math destruction: how big data increases inequality and threatens democracy. New York: Crown, Penguin Random House; 2016.
174. Szeffler SJ. Asthma across the lifespan: time for a paradigm shift. *J Allergy Clin Immunol* 2018;142:773–80. [PubMed: 29627424]
175. Martinez-Castaldi C, Silverstein M, Bauchner H. Child versus adult research: the gap in high-quality study design. *Pediatrics* 2008;122:52–7. [PubMed: 18595986]

Figure.

Machine learning is most often categorized into supervised, unsupervised, and reinforcement learning. Deep learning can be either supervised or unsupervised, but is shown as its own category rather than having 2 subcategories. Transfer learning is a unique application of machine learning that could have any underlying learning model structure. Peripheral nodes show examples (not comprehensive) of research and clinical tasks for which each type of learning might be appropriate.