APS European Society for Paediatric Research SPR Society for Pediatric Research

REVIEW ARTICLE

Check for updates

Improving child health through Big Data and data science

Zachary A. Vesoulis[1], Ameena N. Husain[1] and F. Sessions Cole[1]✉

Child health is defined by a complex, dynamic network of genetic, cultural, nutritional, infectious, and environmental determinants at distinct, developmentally determined epochs from preconception to adolescence. This network shapes the future of children, susceptibilities to adult diseases, and individual child health outcomes. Evolution selects characteristics during fetal life, infancy, childhood, and adolescence that adapt to predictable and unpredictable exposures/stresses by creating alternative developmental phenotype trajectories. While child health has improved in the United States and globally over the past 30 years, continued improvement requires access to data that fully represent the complexity of these interactions and to new analytic methods. Big Data and innovative data science methods provide tools to integrate multiple data dimensions for description of best clinical, predictive, and preventive practices, for reducing racial disparities in child health outcomes, for inclusion of patient and family input in medical assessments, and for defining individual disease risk, mechanisms, and therapies. However, leveraging these resources will require new strategies that intentionally address institutional, ethical, regulatory, cultural, technical, and systemic barriers as well as developing partnerships with children and families from diverse backgrounds that acknowledge historical sources of mistrust. We highlight existing pediatric Big Data initiatives and identify areas of future research.

Pediatric Research (2023) 93:342–349; https://doi.org/10.1038/s41390-022-02264-9

IMPACT:

- Big Data and data science can improve child health.
- This review highlights the importance for child health of child-specific and life course-based Big Data and data science strategies.
- This review provides recommendations for future pediatric-specific Big Data and data science research.

INTRODUCTION

Despite living in the richest economy in the world, children in the United States have worse health outcomes as assessed by quantitative, standardized metrics than children in other upper income countries.[1] Improvements in child health outcomes over the past 30 years associated with availability of vaccines against common childhood diseases, greater chance of survival after preterm birth, and improved nutrition have not reduced race-, ethnicity-, geography-, and poverty-based disparities in child health.[2,3] Access to large medical, biological, and environmental data sets and progress in data science provide multidimensional data and data analytics that more fully capture the dynamic and complex interactions among genetic background, culture, social determinants of health, development, environment, and biologic risk.[1,4–8] These new tools build upon the progress from systems biology omics-based algorithms in which mechanistic explanations of outcomes were anchored in biologic problems to permit integration of environmental (e.g., pollution), cultural, demographic, and geographic factors that interact with child development and biology to create adverse health outcomes.[6,9,10] For example, machine learning algorithms (approaches that improve automatically with additional data experience), a branch of artificial intelligence,[11] recognize and apply patterns in multidimensional data that train models for prediction and stratification of disease risk based on biologic risk as well as developmental, cultural, and environmental risk determinants of adverse child health outcomes.[5,10,12–14]

The epidemiology of child health (children are generally healthier than adults, represent a smaller fraction of the population (22% <18 years of age in the United States),[15] have a higher risk of rare diseases, and have an increasing prevalence of medical complexity (3–5%),[15] the lack of progress in reducing disparities in child health outcomes, and the difficulties in applying adult-based strategies to improve child health outcomes highlight the potential for Big Data and new data science to impact child health. By combining data from traditional sources, e.g., electronic health records, imaging results, biobanks/registries, and omics measurements, with demographic, cultural, and environmental sources, artificial intelligence-based data science methods may identify previously unrecognized patterns associated with childhood disease without starting with an a priori hypothesis.[6] Recent reports have described combining Big Data from different sources in pediatric oncology, nephrology, and sepsis diagnosis.[16–18] However, these reports underline both the potential and the difficulties of combining data sets with limited data dimensions.[19] For example, in pediatric oncology, combining existing institutional registries improves descriptions of natural history and responses to therapies but does not provide insight

[1]The Edward Mallinckrodt Department of Pediatrics, Washington University in St. Louis School of Medicine, and St. Louis Children's Hospital, St. Louis, MO, USA.
✉email: fcole@wustl.edu

Received: 7 March 2022 Revised: 10 June 2022 Accepted: 28 June 2022
Published online: 16 August 2022

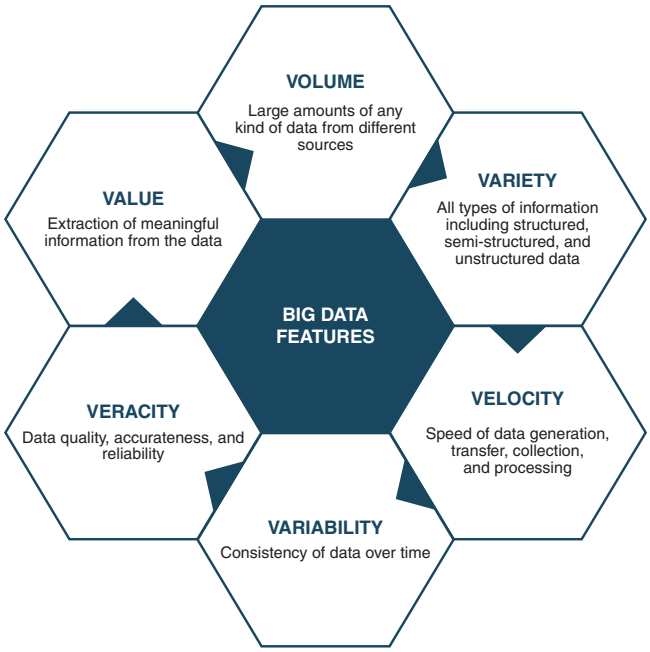

Fig. 1 Big Data features defined by the 6V model. Big Data features defined by the 6V model. Descriptions of each Big Data feature.

into contributions of environment, geography, or other less studied factors to disease risk, response to therapy, or prognosis.[16] Similarly, in pediatric nephrology, development of quantitative definitions of acute kidney injury and descriptions of acute kidney injury epidemiology have not yet fully integrated omics or other potentially informative data elements.[17] Big Data strategies for diagnosis of sepsis in low and middle income countries have helped improve quality of epidemiologic data but have not identified pathophysiologic or structural proposals to improve outcomes.[18] Big Data and data science can accelerate improvements in child health outcomes through reductions in health disparities, improvements in clinical best practices and in quality and safety outcomes, prediction of individual risk of disease progression and response to therapies, increased family and patient-involvement in health care diagnosis and decisions, and discovery of personalized disease mechanisms. We will begin with a brief description of the characteristics of Big Data and data science strategies that will enable improvement in best practices and discovery of disease mechanisms, followed by examples of pediatric Big Data initiatives.

BIG DATA CHARACTERISTICS AND STRATEGIES FOR ANALYSIS

Characteristics

Data production in healthcare is high volume and complex, encompassing a wide variety of inputs and formats including administrative, biomarker (e.g., genomic), physiologic, biometric, laboratory, and imaging. Data may be derived from multiple sources such as electronic health records, medical devices, mobile health platforms, clinical registries, biobanks, and patient self-reports.[5] Within each of these sources, many different data types exist that can be further categorized as structured (e.g., demographics, laboratory results), unstructured (e.g., free text in notes and comments), and semi-structured (a combination of structured and unstructured data).[20]

The fundamental features of Big Data were originally defined in the early 2000s by the 3V model,[21] which includes Volume, Variety, and Velocity. This model has since been extended to 6Vs, adding Variability, Veracity, and Value (Fig. 1).[4,21–25]

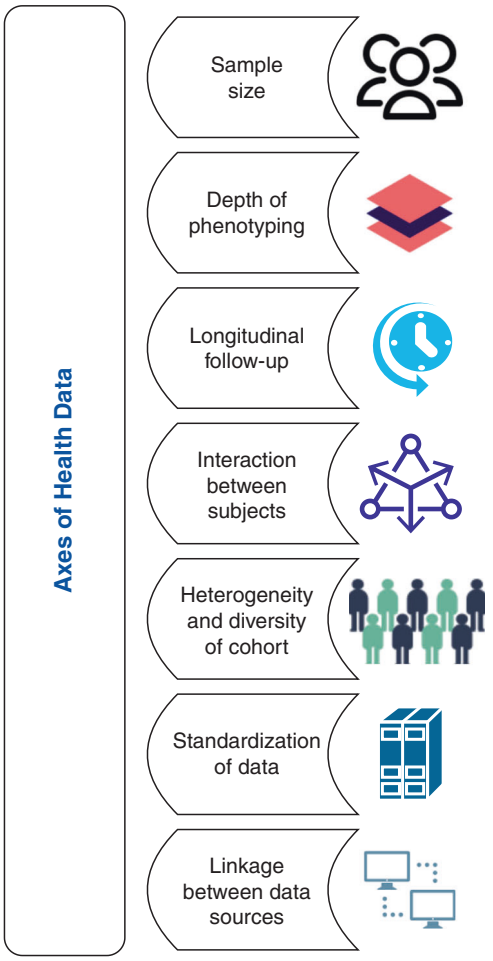

Fig. 2 Quantitative properties represent the complexity of healthcare data. Descriptions of the 7 axes of health data. Adapted from Shilo et al.[5].

With these core features in mind, healthcare data can be further characterized by many quantitative properties that have been previously categorized by Shilo et al. into 7 axes of health data[5] (Fig. 2). "Sample size," a representation of *volume*, is important for achieving sufficient statistical power. However, data with an *n* of 1 (rare diseases) can also provide *value* in the definition of disease trajectory and clinical response.[26] "Depth of phenotyping" describes the *variety* of medical data used to characterize individuals and ranges from the molecular level (e.g., omics, microbiome) to the social level (e.g., demographics, lifestyle, environment, and social determinants of health). Integration of this array of data into a valuable understanding of health can be challenging due to lack of data interoperability and of common definitions. "Longitudinal follow-up" is critical for child health and includes data gathering over different time points, a characterization of *variability*. The Barker hypothesis[27] is a prime example of the importance of life course tracking to improve child and adult outcomes. Similarly, outcomes of underrepresented minorities can be improved with advocacy for engagement and adherence to follow-up in order to obtain complete long-term data.[1] "Interactions between subjects" is an application of *value* to find the connections between subjects (e.g., shared environments, twins), which can increase the statistical power to discover disease mechanisms. "Heterogeneity and diversity" of a cohort to include appropriate representation of the real-world population (factors such as age, sex, ethnicity, socioeconomic status, exposure to different social determinants of health) contribute to the *variety* of

data. "Standardization of data" and "linkage between data sources" are vital to *veracity* and tackling the challenge of having *variety* in data.

Analysis strategies

As this wide array of data is gathered, finding meaningful results requires application of appropriate analytic strategies, the basis of data science. Three overarching approaches to analysis have been used: descriptive, predictive, and counterfactual.[28,29] Descriptive analysis utilizes conventional parametric and non-parametric statistics to provide quantitative estimates of central tendency and variability. Strengths of this approach include the ability to condense large amounts of data into single summative metrics and a high degree of explainability. It is the most common form of analysis for real-time and historical data used in clinical research and is frequently the basis for intervention guidelines and protocols.[29] Predictive analysis utilizes observational data to identify relational patterns between variables, either correlation or anti-correlation. Additional variables can be added to the model to control for factors which independently influence outcomes of interest. Once a model has been constructed, it can be "reversed" by entering values for each of the variables to generate a prediction of the probability of the outcome (for binary models) or an estimated value of the outcome (in continuous models).

In contrast to *associations* identified by descriptive and conventional predictive modeling, counterfactual prediction analysis is the foundation of *causal* analysis and inference. In this approach, one starts with the outcome and works backwards to the model inputs to evaluate how changes in the input might reverse or "flip" the outcome. Counterfactual explanations describe the smallest change to the input variables that causes a change from one predicted outcome to another. It is currently the least commonly used analytic approach. However, it has great potential to answer causal questions, for example in the genetic analysis of complex diseases.[30]

Machine learning analysis

Machine learning refers to computer algorithms which utilize artificial neural networks to identify salient variables or "features" from a large pool of candidate variables which, in association, predict outcomes with the greatest accuracy.[31] Although machine learning is widely considered an "advanced" technique, it encompasses a broad array of approaches ranging from simple, conventional regression modeling through deep learning. The strategy for utilizing machine learning is common across all model types; first, the system is "trained" by exposing it to a representative sample cohort of patients with known and labeled outcomes. As the system learns with each pass through the training samples, predictive accuracy improves. Once training has been maximized, the system is validated using a separate cohort of patients never before seen by the model.

As previously noted, "machine learning" is a broad term encompassing many different computational strategies including neural networks, decision trees, support-vector machines, and deep learning. All machine learning is patterned after human neural networks—each variable is conceived as a node where the pathway through the network diverges based on the value at the node before eventual mapping to the outcome at the end of the network. When such a system is "learning," the algorithm makes observations of individual patients and develops a decision tree, a branching graph where the value of a given variable increases or decreases the predicted probability of the outcome and is also influenced by the path through prior branches in the tree.[32,33] As the algorithm is exposed to more examples, the weights of each branch point are iteratively increased or decreased until accuracy can no longer be increased. Not surprisingly, machine learning systems achieve optimal performance when extremely large datasets are available for training.

A significant drawback of machine learning analysis is the challenge of explaining the findings of the algorithm. As opposed to traditional regression modeling, where each variable is actively chosen by the investigator, and the relationship to the outcome is quantified in understandable units, machine learning decisions are made by a computer based on optimizing outcome prediction. In "unsupervised" machine learning, outcome labels are not provided to the computer, and grouping decisions are made without human input. While this approach offers the potential to discover previously unknown connections between variables, it may also result in convoluted and clinically implausible relationships.

BIG DATA CHALLENGES

The rapid growth of Big Data utilization in healthcare has unmasked many limitations and challenges. Data quality, accuracy, completeness, and availability are major hurdles in using large healthcare datasets[4,34] and can lead to inaccurate analysis,[1] biased inference, and false discoveries.[35] Similarly, management and storage of large amounts of data present challenges with maintaining data security and accuracy over time, archiving data, managing data warehouses, and removing/disposing of information. Applying the most appropriate analytic approach to Big Data requires understanding of the technical and quality limitations of Big Data.

Data integration is both a key challenge and a critical component for improving seamless access to robust, reproducible, and diverse sources of data.[36] Sharing data across institutions, for example, can be difficult due to differences in data types, definitions, and formats. Data security and privacy are also concerns when clinical information is shared, restricting the use of patient identifying information. Addressing these issues will require advancements in the standardization of data. The FAIR Guiding Principles (Findability, Accessibility, Interoperability, and Reusability) of data management provide guidelines for data production and data publishing focused on maximizing data quality and usability and are required by the National Institutes of Health for data management and sharing plans for all award applications submitted after January 25, 2023.[37,38] Improvements in data management which incorporate these fundamental principles will be crucial to harnessing the potential of clinical Big Data.

EXAMPLES OF BIG DATA INITIATIVES AND STRATEGIES FOR IMPROVEMENT OF CHILD HEALTH

Although still early in development, pediatric-specific Big Data projects have begun to emerge that include data commons for integration and interrogation of data from inpatient and ambulatory cohorts of children and electronic health record- or vital sign-based data for development of disease risk scores. Here we highlight examples of pediatric Big Data initiatives. Other multi-institutional pediatric data networks are summarized in Table 1.

Genomic Information Commons

Genomic Information Commons (GIC) is a National Institutes of Health (NIH)-funded, multi-institutional effort to provide an extensive, linked database of genotype, phenotype, biospecimens, and electronic health record-derived metadata in a highly accessible, federated database.[39] A cooperative effort that has recently expanded from three to nine academic pediatric centers in the United States, it leverages robust, easy to use computational infrastructure for preliminary data queries, retention and control of all data and biospecimens at member institutions, scalable inclusion of additional member institutions, executed, inter-institutional data use and material transfer agreements, active

Table 1. Pediatric Big Data networks.

Name	Focus	Data sources
Children's Data Network	Linkage and analysis of administrative records across agencies to inform programs and policies	Healthcare data, Social Services, Education
Children's Hospitals Neonatal Database (CHND)	Large valid dataset for level IV NICU patients for comparative clinical outcomes and resource utilization	Periodic EHR extraction into common data model, Children's Hospital Association administrative dataset
Collaborative Pediatric Critical Care Research Network (CPCCRN)	Multi-institutional network for research in pediatric critical care medicine	Research protocols and study results
Genomic Information Commons (GIC)	Linkage of genomic data, phenotypic data, and biospecimen metadata to accelerate discovery and collaboration	EHR, Genomic laboratory results, Research surveys
ImproveCareNow Registry	Centralized data repository of clinical data for children with inflammatory bowel disease (IBD)	Medical record data at time of diagnosis and every outpatient clinic visit for IBD
National COVID Cohort Collaborative (N3C)	Centralized data repository of clinical data for suspected and confirmed COVID-19 patients (all ages)	Periodic EHR extraction into common data model
PCORnet	Partnership of 8 large Clinical Research Networks via a coordinating hub creating a large comprehensive data network to advance research and public health	Periodic EHR extraction into common data model, patient- reported data, and payor data
Pediatric Emergency Care Applied Research Network (PECARN)	Multi-institutional network for research in pediatric emergency medicine	Research protocols and study results
PEDSnet	Pediatric observational research and clinical trials using large comprehensive multi-specialty network (member of PCORnet)	Periodic EHR extraction into common data model
PhysioNet	Free access to large collections of physiological and clinical data and related open-source software	Standardized data repositories
TriNetX	Large international network and data repository with web-based platform to explore data for research, protocol design, cohort identification, and real-world data analysis	Direct links to health care organizations with specific data repositories, and periodic EHR extraction into common data model
Vermont Oxford Network (VON)	Data repository for very low birth weight infants and all NICU admissions to advance quality improvement, research, and education	Periodic EHR extraction into common data model

participation by patients and families in defining network operations and research priorities, and large and diverse patient populations for genomic discovery and identification of potential therapies. Considerable effort has been made to maximize the value and usability of the data for investigators while maintaining high privacy and security standards.

PEDSnet

PEDSnet is a national pediatric learning health system with a multi-specialty network of collaborators from Children's Hospitals across the United States.[40] Using a common data model from its original funding source, the Patient-Centered Outcomes Research Institute (PCORI), this network formed a centralized data sharing environment with executed, institutional data use agreements to create large datasets of pediatric clinical data extracted from electronic health records which enable communities of patients and clinicians to perform research and quality improvement projects that improve child health.[40,41] PEDSnet currently includes longitudinal clinical data from 2009 for over 6.5 million children, about 9% of all the children in the US.[42]

Clinical data from electronic health records are gathered in quarterly cycles from the contributing sites in a structured and templated manner. The PCORI Common Data Model ensures terminology and data details are standardized and permits interoperability with other PCORI-sponsored Clinical Data Research Networks. An extensive data quality assessment process is used for careful analysis of the quality and characteristics of the data from each participating site.[43] Data issues are categorized,[44,45] reviewed by data scientists, and discussed with each submitting site for resolution to address four dimensions of data quality, fidelity, consistency, accuracy, and completeness.[43]

Data security and privacy are addressed with several steps that include storage of limited datasets in the PEDSnet Data Coordinating Center without patient identifiers. When data are requested by researchers, the minimum necessary aggregated data are provided, and institution-specific information is combined into "counts." PEDSnet data have been used in 61 publications (June, 2022) that include disease-specific, quality and safety, and coronavirus disease 2019 (COVID-19)-related questions.[46–49]

PhysioNet

PhysioNet is a large collection of clinical and physiologic data from both inpatient (e.g., traumatic brain injury) and ambulatory (e.g., gait analysis) venues that includes open-source tools for computational analysis.[50] Initially established in 1999 with NIH support, it is now structured into three components:

1. "PhysioBank"—an extensive archive of digitized physiologic signals from fetal, pediatric, and adult sources.
2. "PhysioToolkit"—an open-source collection of tools for the processing and analysis of physiologic signals.
3. Extensive documentation and tutorials for new and advanced users.

Of the 202 PhysioNet databases, 12 are fetal or pediatric specific. Although most database access is free and without restriction, a subset of databases requires registration and completion of a data use agreement. In addition to providing a data repository that is compliant with the FAIR guidelines, PhysioNet has helped to standardize multiple types of data file formats.[50–52] This standardization expands the number of software

tools that can be used to conduct analyses by researchers including many free and open-source options, providing equitable data access for researchers with limited resources such as those in low and middle-income countries.

Electronic Health Record-based risk scores

Late onset neonatal sepsis is a significant contributor to morbidity and mortality. The nSOFA score[53] is a sepsis prediction score, based upon the adult SOFA (Sequential Organ Failure Assessment) score and pSOFA (a pediatric variation developed for older children[54]) score, which utilizes elements from electronic health records to identify infants at high risk for sepsis related mortality. As with the adult SOFA score, a rise in the nSOFA score was highly correlated with sepsis related mortality—a difference that could be detected within 6–12 h of sepsis evaluation. This score has since been validated in a multicenter cohort of more than 600 infants[55] with excellent performance, noting an area under the curve (AUC) of 0.88 for the prediction of mortality. The nSOFA score has also been shown to discriminate between survival and non-survival on the first day of life in extremely preterm infants.[56,57] These examples highlight the potential for application of data science analytics on data extracted from electronic health records to generate useful tools for severity of illness stratification and targeted treatments.

In older, hospitalized children, identification and stratification of illness severity and need for critical care have used tools initially developed and validated in adult populations[58] (e.g., Early Warning Scores (EWS) such as the National Early Warning Score (NEWS)).[59,60] In a retrospective study, 2–3% of pediatric hospital admissions experience cardiopulmonary arrest and require resuscitation.[61] The Pediatric Early Warning System (PEWS) score[62] provides a similar predictive model for the pediatric inpatient population. In subsequent validation, PEWS scores identified patients at risk of deterioration 12 h in advance of clinically apparent deterioration,[63] reduced the risk of emergency response calls shortly after admission from the Emergency Department,[64] improved timely and orderly transfer of patients to the ICU,[65] and increased the number of days without medical codes outside the ICU.[66]

Physiology-based risk scores

Over the last two decades in the NICU, the use of multiple devices for monitoring physiology-based signals including the electrocardiogram (ECG), pulse oximetry, respiratory rate, arterial blood pressure, transcutaneous partial pressure of carbon dioxide (CO_2), cerebral and organ oximetry (via near-infrared spectroscopy (NIRS)), and electroencephalogram (EEG) has increased. Despite this broad array of available data, clinicians use these physiologic biomarker data almost exclusively for in-the-moment decisions and most often from only one signal, such as targeted oxygen saturation thresholds to reduce risk of retinopathy of prematurity (ROP). Computational integration of individual or multiple monitored physiologic biomarkers over time may reveal unrecognized patterns of pathology.

For example, the Heart Rate Observation (HeRO) score uses ECG characteristics,[67] comprised of beat-to-beat variability in heart rate, accelerations, and decelerations (indicative of autonomic nervous system tone), to predict continuous sepsis risk in the next 24 h.[68] Extensive neonatal validation testing has demonstrated the HeRO score's superior performance over more traditional laboratory or clinical assessments and can reduce sepsis-related mortality by as much as 20%.[69,70]

Using a similar approach to the HeRO score, several groups have used other quantitative physiologic biomarkers (e.g., heart rate variability) to predict adverse long term neurodevelopmental outcomes,[71–73] moderate-severe MRI abnormalities, or death.[72,74] Similarly, quantitative characteristics of continuous EEG monitoring can be used to predict the later occurrence of seizures[75] and outcomes at 24 months[76] and 5 years.[77]

RESEARCH GAPS AND FUTURE DIRECTIONS

Despite the potential of Big Data strategies to improve child health and patient safety by trans-institutional identification of rare patient phenotypes, adverse patient events (e.g., sentinel pediatric events or adverse drug reactions),[78] and responses to therapies, significant barriers remain in the practical application of research strategies to real-world bedside care.[79–81] One of the most significant barriers is the lack of a universal, interoperable, modular system for capturing and sharing medical data. Proprietary and institutionally cloistered electronic health record systems limit discovery of critical components of best clinical and nursing practices[82] and of pediatric-and disease-specific patient characteristics, disease risk, and adverse events. In addition, the large volume of data generated during healthcare delivery requires intentional system design that optimizes future data usability and minimizes cost for data extraction, reformatting, and loss. Similarly, although linkage of individual medical records longitudinally across maternal/fetal, neonatal, child, and adult epochs would be of great value for discovery of fetal, neonatal, and childhood origins of pediatric and adult diseases, such a system remains largely impracticable unless all the care for an individual is obtained within a single health system across the life course. Even within open source electronic health record systems, multiple data formats which vary according to system and region reduce interoperability across institutions.[83] The PRISM model described by Hirschfeld et al.[84] provides a theoretical framework for the future of intentional system design that captures elements of the health phenotype across four dimensions (experience, performance, adaptability, potential) and results in a life course model of an individual's health (an Ideal Health Prism) which could enable comprehensive study of the fetal and childhood origins of childhood and adult diseases.

Through several different programs, such as the Big Data to Knowledge (BD2K) program, the NIH has emphasized the importance of using FAIR principles to insure availability of NIH-funded project data for the scientific community.[37] Through resources such as GIC, PhysioNet, and PEDSNet, secure pediatric data with common formatting models are becoming available. However, compliance with Health Insurance Portability and Accountability Act of 1996 (HIPAA)-associated privacy rules often necessitates extensive manual data review and modification to ensure that all protected health information has been removed. For example, removal or shifting of all elements of date and time (high value Big Data components) in a truly random fashion and uniformly across elements for linkage consistency due to HIPAA protection is both laborious and prone to introduce errors.

Another challenge is the integration of multimodal data. Most of the analytic strategies previously described utilized data from a single source (e.g., the electronic health record) and along a similar time scale. Although this approach simplifies the data collection process, such "siloed" data provide an incomplete understanding of the disease process. Several recent projects[85–87] have demonstrated that building a complex model using multimodal data with different scales can generate neonatal outcome predictions with greater accuracy than single-domain predictions alone. Examples of different scales include race or genetic background, which are immutable characteristics, sepsis status which is discrete but may evolve over time, and vital signs, such as heart rate or blood pressure, which are continuously changing. Further development of artificial intelligence-based data science tools which integrate genomic susceptibility with developmental epoch, environmental factors, social determinants of health, maternal/fetal characteristics, and family/patient self-reported data will be necessary.[88] Meta-dimensional analysis, specifically

concatenation-based integration, is one potential strategy but is not yet in routine use.[89]

Life course research is an example of integration of multiple data types from diverse sources (e.g., institutional data warehouses and research repositories) to capture the complexity of health trajectories.[14,90,91] Incorporation of geocoded data and environmental factors as well as patient-reported measures such as social well-being into the electronic health record represent concrete strategies for greater inclusion and more accurate representation of populations currently underrepresented in research and permit analysis of the impact of social determinants of health on disease pathogenesis and response to therapies.[91]

A significant note of caution must be raised about racial bias in the use of physiology and electronic health record-based Big Data. Two sources of error may contribute to data-related bias. First, devices may not reliably capture measures owing to differences in physiology or phenotype. For example, significant recent attention has focused on the poor performance of pulse oximetry in adult and neonatal African-American patients.[92,93] Lack of inclusion of melanin's light absorption in the red and infrared spectrum in the underlying pulse oximeter algorithm increases the risk of occult hypoxemia in both adult and neonatal populations[94] and of adverse neonatal outcomes.[95,96] Second, even when collected data are reliable, machine learning models must be trained on representative samples that avoid racial bias. As recently demonstrated in a comparison of a new, intentionally designed machine learning algorithm for the prediction of ICU mortality with several widely used scores (APACHE, SAPS II, and MEWS), at least two of these systems (SAPS II and MEWS) were found to have significant racial bias.[97] As with occult hypoxemia, the potential risk of harm comes from false negatives or underestimated disease risk. The equal opportunity difference analysis performed in this study is an optimal tool to identify these deficiencies.

The longstanding problem of gaps in studies of medications and medical devices in pediatric age groups represents an important priority for Big Data and data science to improve child health. As a consequence of these gaps, between 25% and 90% of medications are prescribed to children in an "off label" manner without regulatory approval.[98] Instead of data-driven use, treatment options expand organically through extrapolation of adult data,[99] anecdotal reports by providers, and practice drift.[100,101] Although the primary focus of Big Data and data science has been on improving diagnosis of disease, elucidating mechanisms, and predicting outcomes, these same data science tools can and should be used to identify treatment response in children from real world data. For example, multicenter, federated data commons and advanced data analytics can be leveraged to identify and pool small numbers of infants and children at individual institutions into sample sizes which permit statistically valid examination of treatment response and adverse outcomes. Recently, real world data from electronic health records and other sources have been successfully analyzed to obtain regulatory approval for previously off-label medications in children.[102]

SUMMARY

The urgency of the COVID-19 pandemic has demonstrated the potential for rapid application of Big Data and data science to integrate and analyze electronic health record data across health care systems and countries for identification of child-specific disease characteristics, best clinical practices, and responses to therapeutic interventions.[103,104] These studies suggest the feasibility of the application of Big Data and data science to child health questions and the potential impact of such studies on prediction and mitigation of disease risk over decades of life. Realizing the potential of these tools for integrating genetic risk with developmental epoch, environmental factors, social determinants of health, patient- and family-reported data, and disease biology will require funding prioritization from the NIH and other agencies, unprecedented collaboration among institutions, investigators, and patients/families, consolidation of existing data networks, and child health-specific innovation.

DATA AVAILABILITY

All data generated or analyzed for this review are included in this published article.

REFERENCES

1. Perez, L. G., Peet, E. D., Vegetabile, B. & Shih, R. A. Big Data needs and challenges to advance research on racial and ethnic inequities in maternal and child health. *Womens Health Issues* **32**, 90–94 (2022).
2. GBD 2017 Child and Adolescent Health Collaborators & Reiner, R. C. et al. Diseases, injuries, and risk factors in child and adolescent health, 1990 to 2017: findings from the Global Burden of Diseases, Injuries, and Risk Factors 2017 Study. *JAMA Pediatr.* **173**, e190337 (2019).
3. Ely, D. M. & Driscoll, A. K. Infant mortality in the United States, 2017: data from the period linked birth/infant death file. *Natl Vital Stat. Rep.* **68**, 1–20 (2019).
4. Pablo, R. G. J. et al. Big data in the healthcare system: a synergy with artificial intelligence and blockchain technology. *J. Integr. Bioinform.* **19**, 20200035 (2021).
5. Shilo, S., Rossman, H. & Segal, E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat. Med.* **26**, 29–38 (2020).
6. Hulsen, T. et al. From big data to precision medicine. *Front. Med.* **6**, 34 (2019).
7. U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion. Healthy people 2030. https://health.gov/healthypeople/objectives-and-data/social-determinants-health (2022).
8. Bennett, T. D. et al. Data science for child health. *J. Pediatr.* **208**, 12–22 (2019).
9. Obermeyer, Z. & Emanuel, E. J. Predicting the future - big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **375**, 1216–1219 (2016).
10. MacEachern, S. J. & Forkert, N. D. Machine learning for precision medicine. *Genome* **64**, 416–425 (2021).
11. Hunt, X. et al. Artificial intelligence, big data, and mhealth: the frontiers of the prevention of violence against children. *Front. Artif. Intell.* **3**, 543305 (2020).
12. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-generation machine learning for biological networks. *Cell* **173**, 1581–1592 (2018).
13. Lu, C. Y., McMahon, P. M. & Wu, A. C. Modeling genomic screening in newborns. *JAMA Pediatr*. https://doi.org/10.1001/jamapediatrics.2021.5798 (2022).
14. Goulooze, S. C. et al. Beyond the randomized clinical trial: innovative data science to close the pediatric evidence gap. *Clin. Pharm. Ther.* **107**, 786–795 (2020).
15. Boat, T. F. & Whitsett, J. A. How can the pediatric community enhance funding for child health research. *JAMA Pediatr.* **175**, 1212–1214 (2021).
16. Major, A., Cox, S. M. & Volchenboum, S. L. Using big data in pediatric oncology: current applications and future directions. *Semin. Oncol.* **47**, 56–64 (2020).
17. Sutherland, S. M. Big data and pediatric acute kidney injury: the promise of electronic health record systems. *Front. Pediatr.* **7**, 536 (2019).
18. Iregbu, K. et al. Global health systems' data science approach for precision diagnosis of sepsis in early life. *Lancet Infect. Dis.* **22**, e143–e152 (2021).
19. Martínez-García, M. & Hernández-Lemus, E. Data integration challenges for machine learning in precision medicine. *Front. Med.* **8**, 784455 (2022).
20. AnalytixLabs. Characteristics of Big Data. A complete guide. Blogs & updates on data science, business analytics, AI machine learning. https://www.analytixlabs.co.in/blog/characteristics-of-big-data/ (2021).
21. Bello-Orgaz, G., Jung, J. J. & Camacho, D. Social big data: recent achievements and new challenges. *Int. J. Inf. Fusion* **28**, 45 (2016).
22. Ishawarappa & Anuradha, J. A brief introduction on Big Data 5Vs characteristics and Hadoop technology. https://cyberleninka.org/article/n/1071853/viewer (2015).
23. SearchDataManagement. The 5 V's of Big Data. https://searchdatamanagement.techtarget.com/definition/5-Vs-of-big-data (2022).
24. Luo, J., Wu, M., Gopukumar, D. & Zhao, Y. Big Data application in biomedical research and health care: a literature review. *Biomed. Inf. Insights* **8**, 1–10 (2016).
25. Andreu-Perez, J., Poon, C. C. Y., Merrifield, R. D., Wong, S. T. C. & Yang, G. Z. Big Data for health. *IEEE J. Biomed. Health Inf.* **19**, 1193–1208 (2015).
26. Brokamp, E. et al. One is the loneliest number: genotypic matchmaking using the electronic health record. *Genet. Med.* **23**, 1830–1832 (2021).
27. Kwon, E. J. & Kim, Y. J. What is fetal programming?: a lifetime health is under the control of in utero health. *Obstet. Gynecol. Sci.* **60**, 506–519 (2017).
28. Hernán, M. A., Hsu, J. & Healy, B. A second chance to get causal inference right: a classification of data science tasks. *CHANCE* **32**, 42–49 (2019).

29. ProjectPro. Types of analytics: descriptive, predictive, prescriptive analytics. https://www.projectpro.io/article/types-of-analytics-descriptive-predictive-prescriptive-analytics/209 (2022).
30. Hu, P., Jiao, R., Jin, L. & Xiong, M. Application of causal inference to genomic analysis: advances in methodology. *Front. Genet*. **9**, 238 (2018).
31. Shalev-Shwartz, S. & Ben-David. S. *Understanding Machine Learning: From Theory to Algorithms* (Cambridge University Press, 2014).
32. Chollet, F. *Deep Learning with Python* (Manning Publications Co., 2018).
33. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
34. Sanders, C. et al. Understanding the limits of large datasets. *J. Cancer Educ.* **27**, 664–669 (2012).
35. Wang, W. & Krishnan, E. Big Data and clinicians: a review on the state of the science. *JMIR Med. Inf.* **2**, e1 (2014).
36. Sinha, A., Hripcsak, G. & Markatou, M. Large datasets in biomedicine: a discussion of salient analytic issues. *J. Am. Med. Inf. Assoc.* **16**, 759–767 (2009).
37. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
38. NIH. NOT-OD-21-013: Final NIH policy for data management and sharing. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html (2022).
39. Mandl, K. D. et al. The Genomics Research and Innovation Network: creating an interoperable, federated, genomics learning system. *Genet. Med.* **22**, 371–380 (2020).
40. Forrest, C. B. et al. PEDSnet: a national pediatric learning health system. *J. Am. Med. Inf. Assoc.* **21**, 602–606 (2014).
41. Forrest, C. B., Margolis, P., Seid, M. & Colletti, R. B. PEDSnet: how a prototype pediatric learning health system is being expanded into a national network. *Health Aff.* **33**, 1171–1177 (2014).
42. PEDSnet. Home. http://pedsnet.org (2022).
43. PEDSnet. PEDSnet data quality. http://pedsnet.org (2022).
44. Khare, R. et al. Predicting causes of data quality issues in a clinical data research. *Netw. AMIA Summits Transl. Sci. Proc.* **2018**, 113–121 (2018).
45. Khare, R. et al. A longitudinal analysis of data quality in a large pediatric data research network. *J. Am. Med. Inf. Assoc.* **24**, 1072–1079 (2017).
46. Davis, S. M. et al. Population-based assessment of cardiometabolic-related diagnoses in youth with Klinefelter syndrome: a PEDSnet study. *J. Clin. Endocrinol. Metab*. **107**, e1850–e1859 (2022).
47. Khare, R. et al. Development and evaluation of an EHR-based computable phenotype for identification of pediatric Crohn's disease patients in a National Pediatric Learning Health System. *Learn Health Syst.* **4**, e10243 (2020).
48. Denburg, M. R. et al. Using electronic health record data to rapidly identify children with glomerular disease for clinical research. *J. Am. Soc. Nephrol.* **30**, 2427–2435 (2019).
49. Bailey, L. C. et al. Assessment of 135794 pediatric patients tested for severe acute respiratory syndrome coronavirus 2 across the United States. *JAMA Pediatr.* **175**, 176–184 (2021).
50. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, E215–E220. https://doi.org/10.1161/01.CIR.101.23.e215 (2000).
51. Kemp, B. & Olivan, J. European data format "plus" (EDF+), an EDF alike standard format for the exchange of physiological data. *Clin. Neurophysiol. J.* **114**, 1755–1761 (2003).
52. Shafranovich, Y. Common format and MIME type for comma-separated values (CSV) files. RFC Editor. https://www.rfc-editor.org/rfc/rfc4180.txt (2005).
53. Wynn, J. L. & Polin, R. A. A neonatal sequential organ failure assessment score predicts mortality to late-onset sepsis in preterm very low birth weight infants. *Pediatr. Res.* **88**, 85–90 (2020).
54. Matics, T. J. & Sanchez-Pinto, L. N. Adaptation and validation of a pediatric sequential organ failure assessment score and evaluation of the sepsis-3 definitions in critically ill children. *JAMA Pediatr.* **171**, e172352 (2017).
55. Fleiss, N. et al. Evaluation of the neonatal sequential organ failure assessment and mortality risk in preterm infants with late-onset infection. *JAMA Netw. Open* **4**, e2036518 (2021).
56. Travers, C. P., Carlo, W. A. & Ambalavanan, N. The future of outcome prediction for preterm infants in the neonatal ICU. *Am. J. Respir. Crit. Care Med.* **205**, 6–8 (2022).
57. Lavilla, O. C. et al. Hourly kinetics of critical organ dysfunction in extremely preterm infants. *Am. J. Respir. Crit. Care Med.* **205**, 75–87 (2022).
58. Morgan, R., Williams, F. & Wright, M. An early warning scoring system for detecting developing critical illness. *Clin. Intensive Care* **8**, 100 (1997).
59. Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E. & Featherstone, P. I. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* **84**, 465–470 (2013).
60. Alam, N. et al. The impact of the use of the Early Warning Score (EWS) on patient outcomes: a systematic review. *Resuscitation* **85**, 587–594 (2014).
61. Reis, A. G., Nadkarni, V., Perondi, M. B., Grisi, S. & Berg, R. A. A prospective investigation into the epidemiology of in-hospital pediatric cardiopulmonary resuscitation using the international Utstein reporting style. *Pediatrics* **109**, 200–209 (2002).
62. Duncan, H., Hutchison, J. & Parshuram, C. S. The Pediatric Early Warning System score: a severity of illness score to predict urgent medical need in hospitalized children. *J. Crit. Care* **21**, 271–278 (2006).
63. Akre, M. et al. Sensitivity of the pediatric early warning score to identify patient deterioration. *Pediatrics* **125**, e763–e769 (2010).
64. Frascogna, M. N., Merkle, E., Dowdy, K. & Seals, S. The effect of pediatric early warning score use on emergency response calls after admission from the pediatric emergency department. *Pediatr. Emerg. Care* **37**, e930–e933 (2021).
65. Agulnik, A. et al. Impact of implementing a Pediatric Early Warning System (PEWS) in a pediatric oncology hospital. *Pediatr. Qual. Saf.* **3**, e065 (2018).
66. Demmel, K. M., Williams, L. & Flesch, L. Implementation of the Pediatric Early Warning Scoring System on a Pediatric Hematology/Oncology Unit. *J. Pediatr. Oncol. Nurs.* **27**, 229–240 (2010).
67. Fairchild, K. D. & O'Shea, T. M. Heart rate characteristics: physiomarkers for detection of late-onset neonatal sepsis. *Clin. Perinatol.* **37**, 581–598 (2010).
68. Griffin, M. P. et al. Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. *Pediatr. Res.* **53**, 920–926 (2003).
69. Moorman, J. R. et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial. *J. Pediatr.* **159**, 900.e1–906.e1 (2011).
70. Fairchild, K. D. et al. Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. *Pediatr. Res.* **74**, 570–575 (2013).
71. Goulding, R. M. et al. Heart rate variability in hypoxic ischemic encephalopathy: correlation with EEG grade and 2-y neurodevelopmental outcome. *Pediatr. Res.* **77**, 681–687 (2015).
72. Massaro, A. N. et al. Heart rate variability in encephalopathic newborns during and after therapeutic hypothermia. *J. Perinatol.* **34**, 836–841 (2014).
73. Vesoulis, Z. A., Rao, R., Trivedi, S. B. & Mathur, A. M. The effect of therapeutic hypothermia on heart rate variability. *J. Perinatol.* **37**, 679–683 (2017).
74. Vergales, B. D. et al. Depressed heart rate variability is associated with abnormal EEG, MRI, and death in neonates with hypoxic ischemic encephalopathy. *Am. J. Perinatol.* **31**, 855–862 (2014).
75. Jain, S. V., Zempel, J. M., Srinivasakumar, P., Wallendorf, M. & Mathur, A. Early EEG power predicts MRI injury in infants with hypoxic-ischemic encephalopathy. *J. Perinatol.* **37**, 541–546 (2017).
76. Murray, D. M., Boylan, G. B., Ryan, C. A. & Connolly, S. Early EEG findings in hypoxic-ischemic encephalopathy predict outcomes at 2 years. *Pediatrics* **124**, e459–e467 (2009).
77. Murray, D. M., O'Connor, C. M., Ryan, C. A., Korotchikova, I. & Boylan, G. B. Early EEG grade and outcome at 5 years after mild neonatal hypoxic ischemic encephalopathy. *Pediatrics* **138**, e20160659 (2016).
78. McMahon, A. W. et al. Big Data in the assessment of pediatric medication safety. *Pediatrics* **145**, e20190562 (2020).
79. Hoodbhoy, Z. et al. Machine learning for child and adolescent health: a systematic review. *Pediatrics* **147**, e2020011833 (2021).
80. van de Sande, D. et al. Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. *BMJ Health Care Inf.* **29**, e100495 (2022).
81. Slopen, N. & Heard-Garris, N. Structural racism and pediatric health—a call for research to confront the origins of racial disparities in health. *JAMA Pediatr.* **176**, 13–15 (2022).
82. Cole, F. S. Improving VLBW infant outcomes with big data analytics. *Pediatr. Res.* **90**, 20–21 (2021).
83. Purkayastha, S., Allam, R., Maity, P. & Gichoya, J. W. Comparison of open-source electronic health record systems based on functional and user performance criteria. *Health. Inf. Res.* **25**, 89–98 (2019).
84. Hirschfeld, S. et al. Health measurement model-bringing a life course perspective to health measurement: the PRISM model. *Front. Pediatr.* **9**, 605932 (2021).
85. Temko, A. et al. Multimodal predictor of neurodevelopmental outcome in newborns with hypoxic-ischaemic encephalopathy. *Comput. Biol. Med* **63**, 169–177 (2015).
86. Vesoulis, Z. A., El Ters, N. M., Herco, M., Whitehead, H. V. & Mathur, A. M. A web-based calculator for the prediction of severe neurodevelopmental impairment in preterm infants using clinical and imaging characteristics. *Children* **5**, 151 (2018).
87. Na, J. Y. et al. Artificial intelligence model comparison for risk factor analysis of patent ductus arteriosus in nationwide very low birth weight infants cohort. *Sci. Rep.* **11**, 22353 (2021).
88. Topol, E. J. Individualized medicine from prewomb to tomb. *Cell* **157**, 241–253 (2014).

89. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **16**, 85 (2015).
90. Hanson, H. A. et al. Charting the life course: emerging opportunities to advance scientific approaches using life course research. *J. Clin. Transl. Sci.* **5**, e9 (2020).
91. Hanson, H. A. et al. Opportunities for life course research through the integration of data across Clinical and Translational Research Institutes. *J. Clin. Transl. Sci.* **2**, 156–162 (2018).
92. Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E. & Valley, T. S. Racial bias in pulse oximetry measurement. *N. Engl. J. Med.* **383**, 2477–2478 (2020).
93. Vesoulis, Z., Tims, A., Lodhi, H., Lalos, N. & Whitehead, H. Racial discrepancy in pulse oximeter accuracy in preterm infants. *J. Perinatol.* **42**, 79–85 (2022).
94. Feiner, J. R., Severinghaus, J. W. & Bickler, P. E. Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender. *Anesth. Analg.* **105**, S18–S23 (2007).
95. Vesoulis, Z. A. et al. Early hypoxemia burden is strongly associated with severe intracranial hemorrhage in preterm infants. *J. Perinatol.* **39**, 48–53 (2019).
96. BOOST II United Kingdom Collaborative Group et al.Oxygen saturation and outcomes in preterm infants. *N. Engl. J. Med.* **368**, 2094–2104 (2013).
97. Allen, A. et al. A racially unbiased, machine learning approach to prediction of mortality: algorithm development study. *JMIR Public Health Surveill.* **6**, e22400 (2020).
98. Ristovska, L. Regulations and data sources on pediatric clinical studies in the United States and European Union (White Paper). https://www.nber.org/sites/default/files/2020-08/Regulations%20and%20Data%20Sources%20on%20Pediatric%20Clinical%20Studies%20in%20the%2CUnited%20States%20and%20European%20Union.pdf (2020).
99. Novak, E. & Allen, P. J. Prescribing medications in pediatrics: concerns regarding FDA approval and pharmacokinetics. *Pediatr. Nurs.* **33**, 64–70 (2007).
100. Durrmeyer, X., Vutskits, L., Anand, K. J. S. & Rimensberger, P. C. Use of analgesic and sedative drugs in the NICU: integrating clinical trials and laboratory data. *Pediatr. Res.* **67**, 117–127 (2010).
101. Sharpe, C. et al. Levetiracetam versus phenobarbital for neonatal seizures: a randomized controlled trial. *Pediatrics* **145**, e20193182 (2020).
102. Bolislis, W. R., Fay, M. & Kühler, T. C. Use of real-world data for new drug applications and line extensions. *Clin. Ther.* **42**, 926–938 (2020).
103. Bourgeois, F. T. et al. International analysis of electronic health records of children and youth hospitalized with COVID-19 infection in 6 countries. *JAMA Netw. Open* **4**, e2112596 (2021).
104. Klann, J. G. et al. Validation of an internationally derived patient severity phenotype to support COVID-19 analytics from electronic health record data. *J. Am. Med. Inf. Assoc.* **28**, 1411–1420 (2021).

AUTHOR CONTRIBUTIONS

Z.D.V., A.N.H., and F.S.C. contributed to conception and design of this manuscript; Z.D.V. and A.N.H. drafted the manuscript; Z.D.V., A.N.H., and F.S.C. critically revised the manuscript and reviewed and approved the manuscript for publication.

FUNDING INFORMATION

This work was supported by grants from the National Institutes of Health K23 NS111086 (to Z.A.V.) and the Children's Discovery Institute (F.S.C.).

COMPETING INTERESTS

The authors declare no competing interests.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

No patient consent was required.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to F. Sessions Cole.