

REVIEW ARTICLE



Data harnessing to nurture the human mind for a tailored approach to the child

 Saheli Chatterjee Misra¹✉ and Kaushik Mukhopadhyay²

© The Author(s), under exclusive licence to the International Pediatric Research Foundation, Inc 2022

Big data in pediatrics is an ocean of structured and unstructured data. Big data analysis helps to dive into the ocean of data to filter out information that can guide pediatricians in their decision making, precision diagnosis, and targeted therapy. In addition, big data and its analysis have helped in the surveillance, prevention, and performance of the health system. There has been a considerable amount of work in pediatrics that we have tried to highlight in this review and some of it has been already incorporated into the health system. Work in specialties of pediatrics is still forthcoming with the creation of a common data model and amalgamation of the huge “omics” database. The physicians entrusted with the care of children must be aware of the outcome so that they can play a role to ensure that big data algorithms have a clinically relevant effect in improving the health of their patients. They will apply the outcome of big data and its analysis in patient care through clinical algorithms or with the help of embedded clinical support alerts from the electronic medical records.

Pediatric Research (2023) 93:357–365; <https://doi.org/10.1038/s41390-022-02320-4>

IMPACT:

- Big data in pediatrics include structured, unstructured data, waveform data, biological, and social data.
- Big data analytics has unraveled significant information from these databases.
- This is changing how pediatricians will look at the body of available evidence and translate it into their clinical practice.
- Data harnessed so far is implemented in certain fields while in others it is in the process of development to become a clinical adjunct to the physician.
- Common databases are being prepared for future work.
- Diagnostic and prediction models when incorporated into the health system will guide the pediatrician to a targeted approach to diagnosis and therapy.

The traditional way we look at the vast evidence is gradually expanding at a rate that the human mind finds impossible to process. Hospitals and healthcare-related organizations are generating an enormous amount of data daily. These data sources include structural content like electronic health record (EHR) data on symptoms, vital signs, laboratory data, medication dose, and unstructured data in the form of progress notes, pathology, and radiology reports. Ancillary systems connected to EHRs such as waveforms and pictures are converted into useful data sources. The platform of genomics, proteomics, and metabolomics forms the common source of biological data. Social networks, mobile devices, and self-tracking tools provide sources of rich data. All these data contribute to so-called “big data”.

HOW BIG IS “BIG DATA”?

As per definition, “big data” is a high-volume, high-velocity, and high-variety information asset that demands economically sustainable innovative forms of information processing for better insight and decision making.¹ There is no cut-off dataset file size (like one

gigabyte or one terabyte) beyond which we can label a dataset as big data. Instead, the complexity of the data, which demands advanced technologies to store and analyze, makes it “big data”.² An intricate network of systems with various data engineering skills is required to serve the purpose. One of the most natural approaches is to divide this extensive record into multiple fragments and store them across multiple servers. Hadoop, an open-source platform to store, manage and access data for real-world applications, came into existence in 2011. Since then, it is used by multiple healthcare researchers for big data analytics.^{3–5} One such example is a tool called CHES that stores health-related information in the Hadoop ecosystem. With this tool, Hadoop handles all “big data” issues and users can download a small fraction and perform their data analysis using any statistical software.⁶

BIG DATA ANALYTICS: MACHINE LEARNING AND DEEP LEARNING

The purpose of big data analytics is to extract information and find patterns from this big data available from institutional databases

¹Pediatrics ESIC PGIMS & ESIC Medical College, Diamond Harbour Road, Joka, Kolkata 104 West Bengal, India. ²Pharmacology, AIIMS Kalyani, NH-34Connector, Basantapur, Saguna 741245 West Bengal, India. ✉email: saheli2069@gmail.com

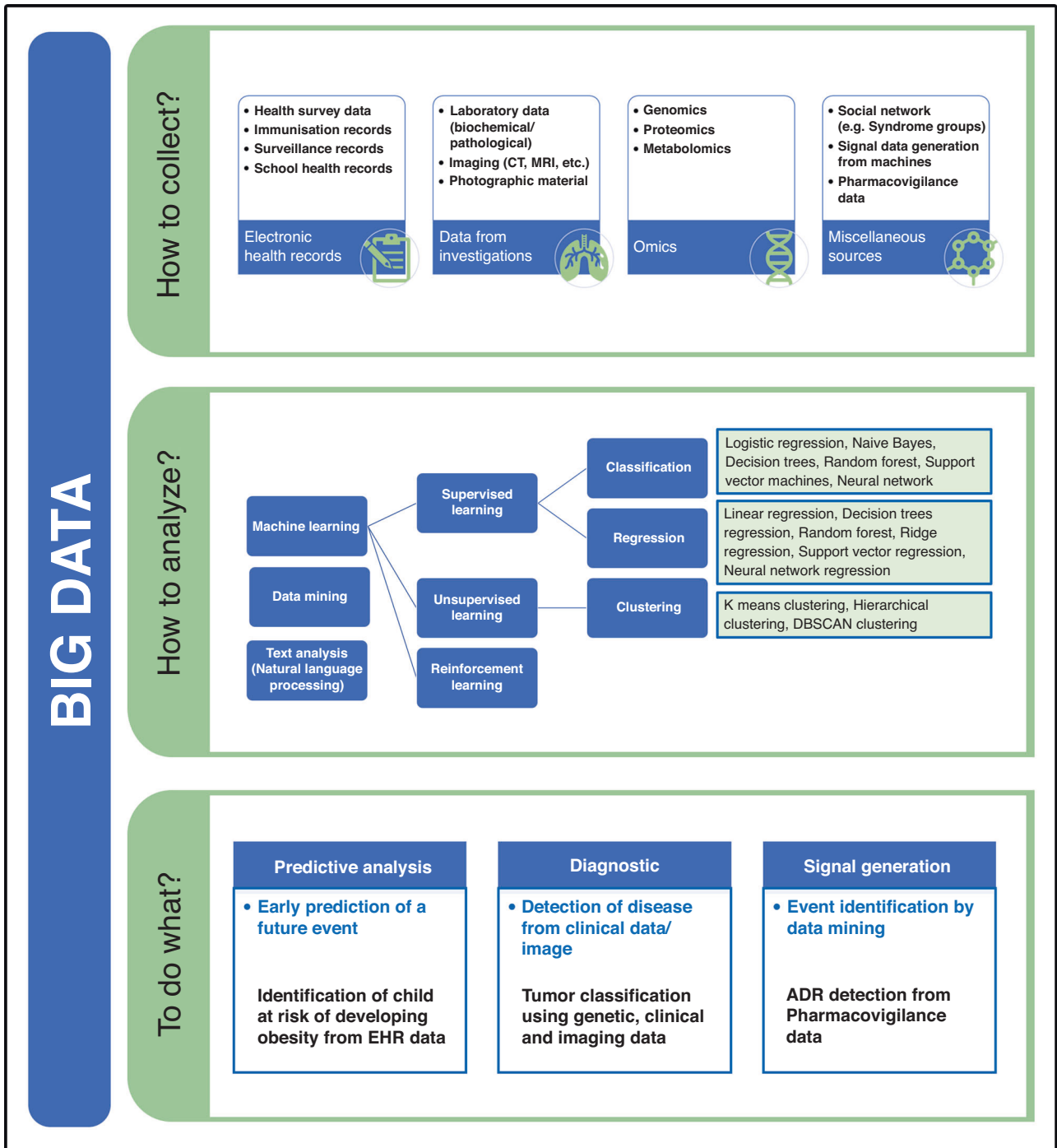


Fig. 1 Schematic representation of Big data. Deals with the various sources,analytics and outcome.

for research to provide better care and outcome. Application of advanced statistical methods like machine learning (ML) and deep learning (DL) has been the key to recent advances in predictive analytics in healthcare. DL is a specialized subfield of ML, a new way to train models using learning successive layers of increasingly meaningful representations.⁷ ML is broadly classified into supervised, unsupervised, and reinforcement learning (RL) (Fig. 1)

In supervised learning, the model is trained with a labeled dataset and known target output. The performance of the model is then validated with the test dataset to check how the model will perform in unknown/new input data. The performance of these models is measured by accuracy, sensitivity, specificity, positive

predictive value, area under curve (AUC) for categorical target variables. For continuous variables, mean squared error, mean absolute error, root mean squared error, etc. have been used. Logistic regression, naive Bayes, random forest, artificial neural network, support vector machines (SVM), and Boosting techniques (AdaBoost, XG Boost) are a few examples of supervised learning models. DL models like convolutional neural network (CNN) and recurrent neural network (RNN) are used to analyze data from images, video, or audio.

Unsupervised learning, on the other hand, does not have any fixed target label as output. Here, only input data is provided, and the model tries to find hidden patterns and draw valuable insights

from the unknown dataset by clustering. K-means clustering, hierarchical clustering, Apriori algorithm are a few examples of unsupervised learning models. RL is a unique ML technique where the model has to make a series of decisions to achieve a goal. If the model makes the right choice, it is rewarded (like +1) and gets a punishment (like -1) for doing wrong. After several similar trial and error experiences, the model gradually learns to finish the task.⁸

BIG DATA IN PEDIATRICS

The handling of data for research in pediatrics started in 1984, but the rapid growth phase began with the dawn of the twenty-first century. The first work of harnessing big data in pediatrics happened when Anderson et al.⁹ in the search for a clue to sudden unexpected infant death, identified maternal smoking as a risk factor. The major areas where big data is now used include predictive analysis, diagnostics, therapy of rare and congenital diseases, pharmacoepidemiology, and pharmacovigilance (Table 1).

DATA HARNESSSED SO FAR

For predictive analysis

Nutrition. Childhood obesity is a growing pandemic that earlier was considered a problem mainly in high-income countries and societies but now also occurring, at a rapid rate, in low-income and middle-income countries.¹⁰ XG Boost model used EHR data up to age 2 years from the Children's Hospital of Philadelphia¹¹ to predict childhood obesity from age >2 to ≤7 years. In time, such models^{11,12} could be used as real-time predictive health trackers to alert clinicians of children at high risk of obesity and improve their decision making. If children with obesity are detected within five years, then lifestyle modification and early screening may help to decrease morbidity and improve the quality of life.

Development. Big data has been able to throw some insight into the management and outcome of children with cerebral palsy (CP).¹³ Though much work has been done in children with autism and attention-deficit hyperactivity disorder^{14,15} to reduce the time to diagnosis, the light at the end of the tunnel is yet to be seen. Large dataset studies have looked at the brain mapping risk for anxiety. Big data analytics may help to unravel the neurobiology of pediatric anxiety.¹⁶ Pruett et al.¹⁷ created a text learning algorithm that has the potential to enable future studies of stuttering using existing EHR data, including investigations into the genetic etiology.

Newborn. There are vast data in the neonatal intensive care unit (ICU) including medical records, physiological data from the signal monitor, and video images of movement. The cause of morbidity and mortality in NICU is primarily centered around premature infants. We know that a lot of financial and human resources are evaporated to reduce the burden of newborn mortality. Toward this end, DL methods are using time-series vital signs data to predict the risk of mortality in preterm infants.¹⁸

There is always the risk of hyaline membrane disease in these infants, which calls for surfactant and ventilation. These preemies often develop bronchopulmonary dysplasia (BPD) that is a preventable disease. Lei et al.¹⁹ predicted from clinical records of 996 premature infants by analyzing 26 variables the risk of developing BPD. They established that the Random forest model (AUC 0.929) could help clinicians predict those preterms at risk of developing the disease to formulate the best treatment plan.

Retinopathy of prematurity (ROP) leads to blindness and is preventable by early detection. A multi-institutional ROP dataset consisting of retinal fundus images and clinical factors was collected as part of the Imaging and Informatics in ROP study.²⁰

The Logistic regression model based on demographics used two predictors, gestational age and visual severity score, which can be collected during a single exam, to identify all subjects who will eventually develop treatment requiring ROP. The study is capable of identifying children who will not develop ROP and help to reduce the unnecessary waiting time for the screening of this disease, let alone intervention.

Available EHR data can identify infants with sepsis in the NICU hours before clinical recognition.²¹ Further prospective trials are required before they can benefit the clinician. Early awareness of sepsis can alert the clinician for judicious medical management.

Pediatric intensive care. At Children's Hospital of Pittsburgh, Pediatric Rothman Index is an analytical tool that is used as an early warning sign for clinical deterioration.²² The Rothman Index graphs can predict gradual health declines that can be missed during handoff between multiple residents and nurses. The software generates a graph displaying the patient's conditions using existing data in the patient's EMR, a live data feed from the patient's continuous monitoring systems, laboratory results, vital signs, and nursing assessments which updates every 1, 5, and 60 min. The analytical tool followed research on a single-center dataset²³ that processed and integrated hundreds of input variables from a patient's electronic medical record to provide an accurate assessment of a child in the ICU. They can predict the patient's risk of desaturation in the next hour²⁴ and may enable early intervention.

Evaluating the factors leading to death in PICU, the common causes were shock, multi-organ failure, septicemia with septic shock, and respiratory failure. Non-invasive ventilation in the form of high-flow nasal cannulae (HFNC) oxygen therapy is nowadays a preferred mode in children. The pediatric intensive care unit (PICU) data were used to make dynamic predictions of HFNC therapy failure within 24 h of initiation²⁵ for escalating respiratory support in critically ill children. Comoretto et al.²⁶ predicted hemodynamic failure development during PICU stay using machine-learning algorithms (Table 1). A similar data source using boosted ensembles of the decision tree has the potential to predict with 75% sensitivity and 70% specificity. They predict pediatric patients' likelihood of developing severe sepsis through automatic monitoring of the patient's EHR.²⁷ These tools will find a place in the clinical decision support, as it is critical to both provide monitoring for this particularly vulnerable population and avoid excessive numbers of false alarms.

Acute kidney injury (AKI) if diagnosed early is now reversible. Various ML algorithms are being developed from readily available EHR variables to predict imminent AKI in hospitalized children accurately.²⁸ PICU data from three centers in the United States and the United Kingdom were used to train an age-dependent ensemble model on pre-disease patterns of physiological measurements to improve the outcome of pediatric AKI by providing early alerting and actionable feedback. This will help in preventing or reducing AKI by implementing early measures such as medication adjustment.²⁹

Emergency and ward. When children present with acute problems in the emergency, triaging these children can be difficult in a busy emergency room. ML-based triage on data from EHR had better discrimination ability to predict clinical outcomes and disposition, with a reduction in under triaging critically ill children and over triaging children who are less ill.^{30,31} Gradient boosted XGB model (AUC 0.82) trained on EHR predicts the probability of ward to ICU transfers more accurately than standard scoring schemes and holds promise for early detection of pediatric inpatients at risk for clinical deterioration in the near future.³²

Pediatric cancer. The Childhood Cancer Initiative has spearheaded efforts toward data sharing to enable the pooling of

Table 1. Studies using big data and big data analytics in different fields of pediatrics.

Author, year	Setting	Medical informatics	Sample size	BDA	Findings
Pang, ¹¹ 2021	Nutrition	EHR	860,510 patients	XG Boost	Childhood obesity model
Pruett, ¹⁷ 2021	Development	Electronic health record	1143 stuttering cases	Text-mining algorithm	Identify stuttering cases in the EHR
Feng, ¹⁸ 2021	NICU	Time-series vital sign data	285 infants	Deep learning	Prediction of preterm mortality
Lei, ¹⁹ 2021	NICU	Clinical records	996 premature infants in the NICU	RF, Boruta	Risk for BPD
Coyner, ²⁰ 2021	NICU	A multi-institutional ROP dataset	Retinal fundus images and clinical factors for 852 subjects	LR	Identify who will eventually develop TR-ROP
Masino, ²¹ 2019	NICU	EHR	Infants hospitalized for 48 h in the neonatal intensive care unit between 2014 and 2017	LR	Identify infants with sepsis in the NICU hours prior to clinical recognition
Aczon, ²³ 2021	PICU	EMR	9070 children admitted to the PICU between 2010 and 2019	Neural network	Accurate continuous assessment for risk of mortality
Ehrlich, ²⁴ 2021	PICU	Virtual PICU database	10,060 children admitted to pediatric care from 2010 to 2020	RNN	Continuous prediction for patient risk of desaturation in the next hour
Pappy, ²⁵ 2021	PICU	Virtual PICU database	All patients admitted to ICU from 2010 to 2020	RNN	Dynamic predictions of HFNC failure within 24 h of initiation
Comoretto, ²⁶ 2021	PICU	Italian Network of PICU registry (TIPNet)	Children admitted to PICU from 2010 to 2020	GLM, RPART, RF, NN, XGB	Haemodynamic failure prediction
Le, ²⁷ 2019	PICU	EHR	9486 pediatric patients 2–17 years, from 2011 to 2016	ML	High-performance severe sepsis detection
Sandokji, ²⁸ 2020	PICU	EMR	8473 encounters from 2014 to 2018	ML	Prediction model of ten EHR variables to predict imminent AKI
Dong, ²⁹ 2021	PICU	EHR	16,863 pediatric critical care patients, 1 month to 21 years	ML	Improve outcome of pediatric AKI
Goto, ³⁰ 2019	Emergency	National hospital Ambulatory Medical care survey	52,037 children <18 years from 2007 to 2015	LR, RF, DNN, Boosted DT	Prediction of clinical outcomes and disposition to help in triaging
Sills, ³¹ 2021	Emergency	Electronic health record data from five EDs in a single healthcare system	9069 ED visits analyzed, 2009–2013	RF, LR	Optimize patient ED flow for pediatric asthma
Mayampurath, ³² 2021	Ward	EHR	38,149 hospital admission at Comer Children's Hospital from 2009 to 2018	RF, XGB, LR	Predicts the probability of ward to ICU transfers
Feng, ³⁵ 2021	Cancer	NCBI Gene Expression Omnibus ArrayExpress	721 microarray samples	DNN	Accuracy of neuroblastoma survival prediction based on gene expression
Kashef, ³⁶ 2020	Cancer	Clinical and medical data	ALL pediatric patients younger than 18 years from 2012 to 2018	XG Boost	Treatment outcome prediction utilizing the SVM algorithm
Khammad, ⁵² 2020	Cancer	MEDLINE database WHO classification of CNS tumors.	100,000 adult and pediatric cases, including rare brain tumor diagnoses	RF, XG Boost	Diagnose primary brain lesions using basic clinical data and histological architecture
Dong, ⁵³ 2021	Cancer	Lab database	1732 bone marrow images	CNN	Diagnosing acute lymphoid leukemia
Porrás, ⁴⁰ 2021	Genetics	Publicly available databases and images collected at the Children's National Hospital	2800 retrospective facial photographs of children, including 1400 children	DNN	Benefit neonatal screening in maternity wards
Bos, ⁴³ 2021	CVS	Mobile ECG (mECG) of genetic heart disease (GHD)-enriched pediatric dataset	Between 2018 and 2020, 686 pediatric patients	DNN/AI	AI-enabled ECG algorithm to determine the QTc
Siontis, ⁴⁴ 2021	CVS	ECG and echocardiogram	318 HCM and 22,996 age- and sex-matched non-HCM control	AI, CNN	Detect pediatric HCM with high accuracy from standard 12-lead ECG.
Mori, ⁴⁵ 2021	CVS	ECG dataset	1192 electrocardiograms of 728 participants from January 2000 to 2017	CNN, LSTM	Accurately diagnose atrial septal defects
Martins, ⁴⁶ 2021	CVS	ECHO dataset	11,646 echocardiography videos from 912 exams, from Brazil and Uganda.	3D CNN	Screening for RHD

Table 1. continued

Author, year	Setting	Medical informatics	Sample size	BDA	Findings
Chen, ⁴⁸ 2020	Radiology	Chest XR database	Radiographs 1–17 years for acute lower airway infection 2018–2019	CNN	Screening tool for common single disease LRTI
Meda, ⁴⁹ 2021	Radiology	Radiology database	104 wrists with rickets and 264 normal wrists radiographic from 2007 to 2017	Object detection model	Identify rickets in <7 years
Smail, ⁵¹ 2020	Radiology	Ultrasound images	2420 sagittal hydronephrosis ultrasound image	CNN	Clinical adjunct for hydronephrosis
Yu, ⁵⁶ 2021	Therapy	Hospital Database	1746 patients aged between 28 days and 18 years from 2013 to 2015	XG Boost RF, catboost, AdaBoost, light GBM	Predict adverse drug events in Chinese pediatric inpatients
Talukdar, ⁵⁹ 2020	Nutrition	Bangladesh Demographic and health survey	6863 children of 0–59 months of age from the 2014 BDHS survey.	RF	Risk of malnutrition
Adegbosin, ⁶⁰ 2020	Nutrition	DHS	1,520,018 children	CNN, hybrid CNN -DNN, LR	Determinants of U5 mortality
Fenta, ⁶¹ 2021	Nutrition	Survey data (2000, 2005, 2011, and 2016) Ethiopia	29,333 children of age 0–59 months	LR, RF, Elastic Net, NN	U5 undernutrition status
Methun, ⁶² 2021	Nutrition	Bangladesh Multi indicator cluster survey	23,099 children aged below 5 years.	LR	Risk factors for U5 morbidity

RMV recurrent neural network, GLM generalized linear models, RPART recursive partition tree, RF random forest, NN neural networks, XGB extreme gradient boosting, catboost Categorical Boosting, AdaBoost Adaptive Boosting, CNN convolutional neural network, LSTM long short-term memory, ARIMA autoregressive integrated moving average, LR logistic regression.

patient-level data for children across the country and the globe.³³ They have combined these resources into Cancer Research Data Commons Ecosystem.³⁴ The neuroblastoma survival prediction³⁵ based on gene expression has helped to provide explanations for a better understanding of the molecular mechanisms underlying neuroblastoma. There is a significant improvement in the treatment outcome prediction of ALL utilizing the SVM algorithm.³⁶ Algorithm developed in genomic data from pediatric AML patients³⁷ have greater precision and accuracy for germline variant detection, and integration of multiple such data will improve risk stratification and therapeutic selection for a better outcome.

For diagnosis

Infectious disease. Infection remains one of the leading causes of mortality in children. Automated diagnosis of childhood pneumonia in resource-constrained settings, compensating for the shortage of expensive equipment and highly trained clinicians, was possible by data mining tools using three vital signs and a biomarker.³⁸ The diagnosis of acute febrile illness and monitoring response to acute severe malaria treatment through ML algorithm based on hematological parameters can be incorporated into clinical support systems, particularly in endemic settings.³⁹

Genetics. Delay in genetic diagnosis is very common, especially in low and middle-income counties. Moreover, cautious clinicians sometimes refer healthy children with atypical facial features for costly genetic evaluation. In a multinational retrospective study, Porras et al.⁴⁰ developed and evaluated a ML-based screening technology using facial photographs with an accuracy of 88% and sensitivity of 90%. As the use of mobile health technology is increasing, as an app, this technology could benefit neonatal screening in maternity wards. In the current era of telemedicine, such a tool would help in remote genetic risk stratification and evaluation of children without requiring them and their families to visit clinical facilities.

Electrocardiogram/echocardiogram. Due to the use of wearables,^{41,42} a considerable amount of data is becoming available for analysis and integration into decision making. Sudden death episodes in children are common and demand screening for both acquired and congenital long QT syndrome. The DNN-enabled QTc algorithm (sensitivity 86%, specificity 94%) in the future through wearable devices may help to detect QT prolongation that predisposes to ventricular arrhythmias and sudden cardiac death in children.⁴³ Hypertrophic cardiomyopathy (HCM) is another cause of sudden death in children. A CNN-based ECG model for HCM detection can predict the disease with high accuracy from the standard 12-lead ECG.⁴⁴ Prediction models⁴⁵ (Table 1) identified atrial septal defects from ECGs that will help predict the disease when the heart murmur remains undetected. Automatic diagnosis of echo-detected rheumatic heart disease (RHD) is feasible with a 3D CNN that has an accuracy of 72.77%.⁴⁶ Further research can potentially reduce the workload of experts, enabling the implementation of more widespread screening programs for RHD worldwide.

Radiology. Artificial intelligence software (i.e., BoneXpertTM v.3.0.3) is used in some radiology departments for the rapid, automated assessment of bone age rather than the traditional and time-consuming manual Greulich–Pyle or Tanner–Whitehouse assessment.⁴⁷ Work with pediatric radiology database shows that models can detect whether a chest X-ray image is abnormal with 92.47% accuracy and identify one of the causes of the lower respiratory tract.⁴⁸ In time, this can also help review the chest X-ray images interpreted by clinicians and help in preventing negligence. This system can be of reasonable diagnostic assistance under limited medical resources. The object detection

model, when trained, will help to identify rickets in children less than 7 years.⁴⁹ Other recent examples of human expert-level performance of high sensitivity and specificity include identifying hemorrhage, mass effect, and hydrocephalus on head computed tomographic (CT) images; colitis on abdominopelvic CT images; and tuberculosis on chest radiographs.⁵⁰ Grading hydronephrosis severity relies on subjective interpretation of renal ultrasound images.⁵¹ It can be a DL-based clinical adjunct for hydronephrosis to the pediatrician, who may benefit in decision making.

Pediatric cancer. Khammad et al.⁵² used decision tree algorithms and neural network on clinical and histological data to design a reliable and interactive computational platform for primary CNS tumor diagnosis. Bone marrow images⁵³ are used to train a CNN, which was effective in classifying WBCs. The performance in diagnosing acute lymphoid leukemia achieved an accuracy of 89% (sensitivity 86%, specificity 95%). The system also performs well at detecting the bone marrow metastasis of lymphoma and neuroblastoma, achieving an average accuracy of 82.93%. Further research is required before this model is implemented in clinical practice.

For pharmacovigilance and pharmacoepidemiology

There are potential applications of big data in pediatric medication safety research in children. Different networks from countries are coming together to combine data from claims, EHR, and surveillance across states. They provide an opportunity to follow children into adulthood and study the effect of medication used in their childhood.

Traditional work with big data from vaccine safety surveillance was used by Yih et al. to show that RotaTeq was associated with excess cases of intussusception in recipients of the first dose. A cohort study using Clinical Practice Research Datalink found that children with epilepsy were at significantly greater risk from medicinal poisoning than those without epilepsy.^{54,55}

Gradient boosted decision tree trained on adverse drug reactions from the EHR of children attending a hospital have predicted risk factors associated with these ADR and may be helpful soon in clinical practice.⁵⁶ Work on pharmacogenomics coupled with EHR for the delivery of precision medicine is ongoing in many hospitals in the US. These results deliver electronic clinical decision support to clinicians when high-risk medications are prescribed for an affected child.⁵⁷ A study by Mohering et al.⁵⁸ identified antimicrobial exposure from EHRs.

For a better health system performance

Big data analysis has been applied to the demographic and health survey data from developing countries to predict risk factors for malnutrition in Bangladesh⁵⁹ and under-five mortality⁶⁰ in 34 low and middle-income countries. These may help to implement preventive health planning for the country. Similarly, the data from the most recent nationally representable cross-sectional Bangladesh Multiple Indicator Cluster Survey predict measures for reducing childhood morbidity in Bangladesh.⁶¹ The retrospective cross-sectional survey data from Ethiopia, a national-representative data, has effectively predicted the under-five undernutrition status in Ethiopian administrative zones⁶² and provided useful information to decision makers trying to reduce child undernutrition.

The ROP dataset study²⁰ shows findings of higher median ROP severity in NCUs without oxygen blenders and pulse oxygenation monitors, which may be utilized for improved access to care for secondary prevention of ROP and may facilitate the assessment of disease epidemiology and NCU resources.

Cerebral palsy (CP) is a broad umbrella that encompasses motor developmental problems due to varied etiology. These children have multiple problems that need a multidisciplinary team. The interaction with this team involves appointment schedules,

outpatient waiting, multiple visits, and travel which becomes a laborious task for parents and caregivers culminating in compromised care for the child. Kurowski et al.⁶³ demonstrated how EHR data helps children with CP interact with a tertiary healthcare system with different specialties. The study may provide a foundation for developing a better system of care that will streamline, cluster services, and reduce frequent acute visits, thus decreasing the burden of care on families and reducing healthcare costs.

Data harnessing in progress for a tailor-made approach in future

Pediatric neurology. A common data model is now available for ML and helps in the binary classification of Interictal epileptiform discharges. Automated seizure detection and forecasting are ongoing before they can be prospectively trialed in children or converted into models.⁶⁴

Newborn. A Harvard necrotizing enterocolitis NEC Database⁶⁵ was developed on patients admitted from 2008 to 2018 to four Harvard-affiliated NICUs with a clinical diagnosis of NEC confirmed by radiograph. Ultra-high accuracy natural language processing (a process of extracting specific information from free-text data entry, such as daily provider progress notes)-based labeling algorithms is identifying radiology reports positive for critical findings of pneumatosis, portal venous gas, and pneumoperitoneum. They may help to predict those infants who will eventually develop NEC using plain films of the abdomen and decrease the high level of morbidity and mortality associated with the disease.

Ulcerative colitis (UC). Advances in next-generation sequencing lead to a greater understanding of the molecular basis of pediatric UC. Multiple modalities such as endoscopy, histology, and imaging ascertain the diagnosis and monitor progress. Large multimodal data analysis can translate established and newly discovered predictive factors into the clinical setting. A clinical decision-support tool, incorporated into the EHR, can be developed to help patient stratification at disease onset and allocate personalized therapies.⁶⁶

Asthma. Asthma is a complex disease associated with multiple risk factors that are difficult to diagnose in children less than 6 years of age. ML models using large complex datasets developed predictive models for childhood asthma.⁶⁷ Using topic modeling of nasal airway expression data, transcriptionally distinct airway endotypes of childhood asthma were identified, corresponding to known clinical phenotypes.⁶⁸ Distinct pathways of disease elucidated by this approach hold promise for the future development of patient-specific therapies.

Pediatric renal disease. Andrea et al.⁶⁹ used the National Patient-Centered Clinical Research Network (PCORnet) Data to develop an algorithm for the identification of primary nephrotic syndromes in children and adults using ICD-9-CM and ICD-10-CM codes. The application of ML may be helpful in identifying risk factors of mortality in dialysis populations once "routine" biomarkers that are cheap and easily available are identified.⁷⁰

Autoimmune diseases. One example where transcriptomics has been highly informative has been in lymphocyte transcriptomics in autoimmunity (SLE, ANCA-associated vasculitis) This has revealed that a pattern of CD8 T-cell exhaustion is predictive of reduced relapses in autoimmunity and is inversely correlated with CD4 co-stimulation. A marker, KAT2B, could potentially have utility as a surrogate marker of clinical outcomes in these diseases. This will enable prognostication of clinical course and identify potential therapeutic targets to reduce relapse in autoimmune diseases.⁷¹

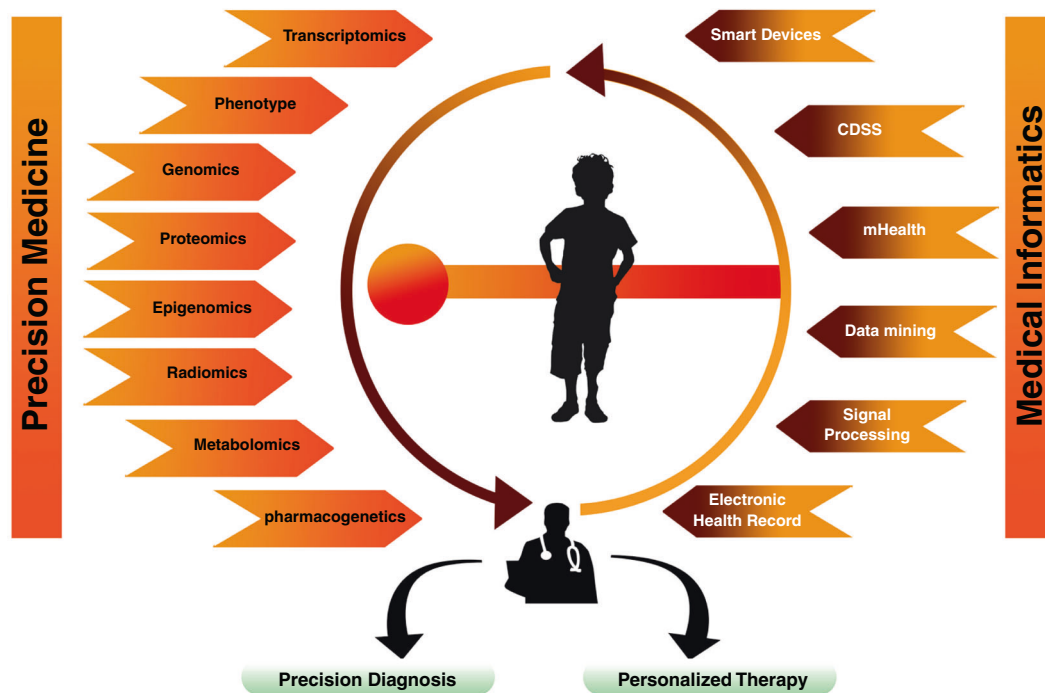


Fig. 2 From Medical informatics to Precision Medicine. World of big data can augment the knowledge and wisdom of clinician for better diagnosis and personalized therapy for the child. CDSS Clinical decision support system.

Type 1 diabetes (T1D). The pathogenesis of T1D in children is still unclear. The intestinal microbiota has shown close relationships with T1D in recent years.⁷² Children at increased genetic risk of type 1 diabetes were prospectively followed from birth in The Environmental Determinants of the Diabetes in the Young (TEDDY) study.⁷³ They conclude that metabolic processes may play a role in triggering islet autoimmunity. Compiling metabolomics and microbiome may help to understand the disease for targeted therapy.

Pitfalls and challenges of big data

EHR, social media, medical transcriptions, and similar unstructured data sources can provide massive chunks of data, but the analytics are complex, costly, time-consuming, requiring skill and effort. Data quality is also an essential challenge in which we need methods for improving data completeness, conformity, and plausibility.⁷⁴ However, access to clinical data is still very restricted due to data sensitivity and ethical issues.

Data standardization or storage in compatible formats with all applications and technologies is another challenge. Some significant challenges that experts deal with during analysis include fragmented heterogeneous data, incompatible formats, unclear transparency during information acquisition and cleansing, and language.⁷⁵ Too much-unrelated data can shift the focus from essential variables and reduce the model speed and performance. Algorithms built based on poor-quality data can lead to misguided decision making by healthcare professionals and policymakers.⁷⁶ External validity and generalizability outside the study population may be an area of concern before implementing a “big data”-based system. A major drawback of some ML and DL models are lack of interpretability by humans because they are “black box” models—only good at discovering the patterns rather than explaining them.⁷⁷ There must be regulations for sharing, de-identifying, securely storing, transmitting, and handling protected health information. This involves privacy laws and legal agreements and requires establishing rules. “Big data” makes the problem of patient privacy protection more significant and more challenging to attain.⁷⁸

Once we overcome these challenges, clear opportunities exist for well-designed diagnosis and decision-support tools to incorporate these data for extracting useful information and providing better outcomes.

PERSPECTIVES AND CONCLUSION

There has been an appreciable amount of research in the healthcare databases that will improve the performance of the health system through proper utilization of resources and judicious use of workforce. Research with data from PICU has progressed at a faster pace helping in clinical decision making and alerting medical personnel. The world of “big data” is growing at a rapid rate. As more genomic data, epigenomic data, methylation array data, transcriptomic data, and proteomic data become available, we can map them to patient phenotype for more personalized care and targeted therapy for children (Fig. 2). Pediatric proteomics, an area of plasma proteomics, is an extremely valuable biological sample to monitor health and disease. In the field of cancer therapy, proteomics has a clear application in the identification of biomarkers for good responders to the treatment of Hodgkin lymphoma and detecting the biochemical signs of high- or low-risk lymphoblastic leukemia patients.⁷⁹ Plasma proteomics has a clear significance in the clinical setting. Advanced research in this field may unravel biomarkers in autoimmunity, blood disease, cardiovascular disease, developmental delay, infection, nutritional imbalance, and even psychiatric conditions. The role of metabolomics in perinatal asphyxia, neonatal nutrition and sepsis, and autism will add to the ocean of omics data.⁸⁰ As whole-genome sequencing data is available, with the advances in technology, genomic diagnosis with a panel-based gene testing will enable rapid and affordable testing of congenital and rare diseases.⁸¹ With a deeper understanding of system biology, there will be an intensive study on the impact of genetic polymorphisms on clinical outcomes and interaction of medications with the microbiome. A large number of pediatric bacterial culture results with genome sequencing data are generated. The application of big data analytics to integrate

routine clinical and laboratory data with omics and sequencing-based outputs may translate personalized medicine initiatives into clinical practice. Predictive modeling can enable clinicians to distinguish children who will benefit from a specific therapeutic intervention from those who will not. Big data analysis and its incorporation into the clinical decision-support system will help us to choose more child-specific management and prevent adverse effects, long-term morbidity, and mortality.

As these changes evolve, the pediatrician will play a different role in ensuring that big data algorithms have a clinically relevant effect on improving the health of their patients. They will be involved in educating their trainees on algorithm development and standardization of data collection through the EMRs. They may even train in analytics to develop strategies for clinical use of big data. Surely with time, most of them will apply the outcome of big data and its analysis in patient care through clinical algorithms or by embedded clinical support alerts from the EMR.

DATA AVAILABILITY

The datasets generated during the current study are included in this published article and available from the corresponding author upon reasonable request. In certain cases, hyperlinks to publicly archived datasets generated during the study are available.

REFERENCES

- Smallwood, C. D. Monitoring big data during mechanical ventilation in the ICU. *Respir. Care* **65**, 894–910 (2020).
- Mashingaidze, K. & Backhouse, J. The relationships between definitions of big data, business intelligence and business analytics. *Int. J. Bus. Inf. Syst.* **26**, 488–505 (2017).
- Belle, A. et al. Big data analytics in healthcare. *Biomed. Res. Int.* **2015**, 370194, <https://doi.org/10.1155/2015/370194> (2015).
- Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. Big data application in biomedical research and health care: a literature review. *Biomed. Inform. Insights* **8**, 1–10. <https://doi.org/10.4137/BII.531559> (2016).
- Harb, H., Mroue, H., Mansour, A., Nasser, A., & Cruz, E. M. A Hadoop-based platform for patient classification and disease diagnosis in healthcare applications. *Sensors (Basel, Switzerland)* **20**, 1931, <https://doi.org/10.3390/S20071931> (2020).
- Batarseh, F. A. & Latif, E. A. Assessing the quality of service using big data analytics: with application to healthcare. *Big Data Res.* **4**, 13–24 (2016).
- Chollet, F. *Deep Learning with Python* 2nd edn 7–8 (Manning Publications Co., 2017).
- Francois-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G. & Pineau, J. An introduction to deep reinforcement learning. *Found. Trends Mach. Learn.* **11**, 219–354 (2018).
- Anderson, T. M. et al. Maternal smoking before and during pregnancy and the risk of sudden unexpected infant death. *Pediatrics* **143**, e20183325, <https://doi.org/10.1542/PEDS.2018-3325> (2019).
- The Lancet Diabetes Endocrinology Childhood obesity: a growing pandemic. *Lancet Diabetes Endocrinol.* **10**, 1 (2022).
- Pang, X., Forrest, C. B., Masino, A. J. & Le-Scherban, F. Prediction of early childhood obesity with machine learning and electronic health record data. *Int. J. Med. Inform.* **150**, 104454, <http://www.elsevier.com/inca/publications/store/5/0/6/0/4/0/> (2021).
- Hammond, R. et al. Predicting childhood obesity using electronic health records and publicly available data. *PLoS One* **14**, e0215571 (2019).
- Fan, H., Li, L., Gilbert, R., O'Callaghan, F. & Wijlaars, L. A machine learning approach to identify cases of cerebral palsy using the UK primary care database. *Lancet* **392**, S33 (2018).
- Bone, D. et al. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *J. Autism Dev. Disord.* **45**, 1121–1136 (2015).
- Bledsoe, J. C. et al. Diagnostic classification of ADHD versus control: support vector machine classification using brief neuropsychological assessment. *J. Atten. Disord.* **24**, 1547–1556 (2020).
- Zacharek, S. J., Kribakaran, S., Kitt, E. R. & Gee, D. G. Leveraging big data to map neurodevelopmental trajectories in pediatric anxiety. *Dev. Cogn. Neurosci.* **50**, 100974 (2021).
- Pruett, D. G. et al. Identifying developmental stuttering and associated comorbidities in electronic health records and creating a phenome risk classifier. *J. Fluency Disord.* **68**, 105847 (2021).
- Feng, J., Lee, J., Vesoulis, Z. A. & Li Fuhai, A. O. Predicting mortality risk for preterm infants using deep learning models with time-series vital sign data. *npj Digit. Med.* **4**, 108, <https://www.nature.com/npjdigitalmed/> (2021).
- Lei, J. et al. Risk identification of bronchopulmonary dysplasia in premature infants based on machine learning. *Front. Pediatr.* **9**, 719352, <https://www.frontiersin.org/journals/pediatrics> (2021).
- Coyner, A. S. et al. A risk model for early detection of treatment requiring retinopathy of prematurity using a deep learning-derived vascular severity score. *Investig. Ophthalmol. Vis. Sci.* **62**, <https://iovs.arvojournals.org/article.aspx?articleid=2773154> (2021).
- Masino, A. J. et al. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. *PLoS One* **14**, e0212665 (2019).
- Suresh, S. Big data and predictive analytics. Applications in the care of children. *Pediatr. Clin. North Am.* **63**, 357–366 (2016).
- Aczon, M. D., Ledbetter, D. R., Laksana, E., Ho, L. V. & Wetzel, R. C. Continuous prediction of mortality in the PICU: a recurrent neural network model in a single-center dataset. *Pediatr. Crit. Care Med.* **22**, 519–529, <http://journals.lww.com/pccmjournal> (2021).
- Ehrlich, L., Ledbetter, D., Aczon, M., Laksana, E. & Wetzel, R. Continuous risk of desaturation within the next hour prediction using a recurrent neural network. *Crit. Care Med.* **49**, 480 (2021).
- Pappy, G., Ledbetter, D., Aczon, M. & Wetzel, R. Early prediction of HFNC failure in the pediatric ICU using a recurrent neural network. *Crit. Care Med.* **49**, 501 (2021).
- Comoretto, R. I. et al. Predicting hemodynamic failure development in PICU using machine learning techniques. *Diagnostics* **11**, 1299 (2021).
- Le, S. et al. Pediatric severe sepsis prediction using machine learning. *Front. Pediatr.* **7**, 413 (2019).
- Sandokji, I. et al. A time-updated, parsimonious model to predict AKI in hospitalized children. *J. Am. Soc. Nephrol.* **31**, 1348–1357 (2020).
- Dong, J. et al. Machine learning model for early prediction of acute kidney injury (AKI) in pediatric critical care. *Crit. Care* **25**, 288 (2021).
- Goto, T., Camargo, C. A., Faridi, M. K., Hasegawa, K. & Freishtat, R. J. Machine learning-based prediction of clinical outcomes for children during emergency department triage. *JAMA Netw. Open* **2**, e186937 (2019).
- Sills, M. R., Ozkaynak, M. & Jang, H. Predicting hospitalization of pediatric asthma patients in emergency departments using machine learning. *Int. J. Med. Inform.* **151**, 104468 (2021).
- Mayampurath, A. et al. Predicting deterioration in hospitalized children using machine learning. *Crit. Care Med.* **49**, 523 (2021).
- Major, A., Cox, S. M. & Volchenbom, S. L. Using big data in pediatric oncology: current applications and future directions. *Semin. Oncol.* **47**, 56–64 (2020).
- NCI Cancer Research Data Commons | CBIIT (accessed 25 February 2022); <https://datascience.cancer.gov/data-commons>.
- Feng, C. et al. A deep-learning model with the attention mechanism could rigorously predict survivals in neuroblastoma. *Front. Oncol.* **11**, 653863 (2021).
- Kashef, A., Khatibi, T. & Mehrvar, A. Treatment outcome classification of pediatric acute lymphoblastic leukemia patients with clinical and medical data using machine learning: a case study at MAHAK hospital. *Inform. Med. Unlocked* **20**, 100399 (2020).
- Alloy, A. P. et al. Novel pediatric AML patient risk stratification by inferred protein activity through integrative network analysis and machine learning. *Cancer Res.* **81**, 1 (2021).
- Naydenova, E., Tsanas, A., Howie, S., Casals-Pascual, C. & De Vos, M. The power of data mining in diagnosis of childhood pneumonia. *J. R. Soc. Interface* **13**, 20160266 (2016).
- Morang'a, C. M. et al. Machine learning approaches classify clinical malaria outcomes based on haematological parameters. *BMC Med.* **18**, 375 (2020).
- Porras, A. R., Rosenbaum, K., Summar, M., Tor-Diez, C. & Lingurar, M. G. Development and evaluation of a machine learning-based point-of-care screening tool for genetic syndromes in children: a multinational retrospective study. *Lancet Digit. Heal.* **3**, e635–e643 (2021).
- Gaffar, S., Gearhart, A. S. & Chang, A. C. The next frontier in pediatric cardiology: artificial intelligence. *Pediatr. Clin. North Am.* **67**, 995–1009 (2020).
- Tandon, A. & De Ferranti, S. D. Wearable biosensors in pediatric cardiovascular disease: promises and pitfalls toward generating actionable insights. *Circulation* **140**, 350–352 (2019).
- Bos, J. M. et al. Artificial intelligence-enabled assessment of the heart rate corrected qt interval using a mobile electrocardiogram device in children and adolescents. *Hear. Rhythm* **18**, S82–S83 (2021).
- Siontis, K. et al. Detection of hypertrophic cardiomyopathy by artificial intelligence-enabled electrocardiography in children and adolescents. *J. Am. Coll. Cardiol.* **77**, 3247 (2021).
- Mori, H. et al. Diagnosing atrial septal defect from electrocardiogram with deep learning. *Pediatr. Cardiol.* **42**, 1379–1387 (2021).

46. Martins, J. F. B. S. et al. Towards automatic diagnosis of rheumatic heart disease on echocardiographic exams through video-based deep learning. *J. Am. Med. Inform. Assoc.* **28**, 1834–1842, <http://jamia.oxfordjournals.org/content/22/e1> (2021).
47. Hwang, S. T., Kang, C. H. & Ahn, K.-S. External validation of X-ray image-based artificial intelligence bone age automatic measurement software for growing children. *Skelet. Radio.* **47**, 1320 (2018).
48. Chen, K. C. et al. Diagnosis of common pulmonary diseases in children by X-ray images and deep learning. *Sci. Rep.* **10**, 17374 (2020).
49. Meda, K. C., Milla, S. S. & Rostad, B. S. Artificial intelligence research within reach: an object detection model to identify rickets on pediatric wrist radiographs. *Pediatr. Radiol.* **51**, 782–791 (2021).
50. Summers, R. M. Deep learning lends a hand to pediatric radiology. *Radiology* **287**, 323 (2018).
51. Smail, L. C., Dhindsa, K., Braga, L. H., Becker, S. & Sonnada, R. R. Using deep learning algorithms to grade hydronephrosis severity: toward a clinical adjunct. *Front. Pediatr.* **8**, 1, <https://www.frontiersin.org/journals/pediatrics> (2020).
52. Khamad, V. et al. Application of machine learning algorithms for the diagnosis of primary brain tumors. *J. Clin. Oncol.* **38**, 2533–2533 (2020).
53. Dong, B. et al. Development and evaluation of a leukemia diagnosis system using deep learning in real clinical scenarios. *Front. Pediatr.* **9**, 693676 (2021).
54. Yih, W. K. et al. Intussusception risk after rotavirus vaccination in U.S. infants. *N. Engl. J. Med.* **370**, 503–512 (2014).
55. Prasad, V., Kendrick, D., Sayal, K., Thomas, S. L. & West, J. Injury among children and young adults with epilepsy. *Pediatrics* **133**, 827–835 (2014).
56. Yu, Z. et al. Predicting adverse drug events in chinese pediatric inpatients with the associated risk factors: a machine learning study. *Front. Pharmacol.* **12**, 659099 (2021).
57. Gregornik, D., Salyakina, D., Brown, M., Roiko, S. & Ramos, K. Pediatric pharmacogenomics: challenges and opportunities: on behalf of the Sanford Children's Genomic Medicine Consortium. *Pharmacogenomics J.* **211**, 8–19 (2020).
58. Moehring, R. W. et al. Development of a machine learning model using electronic health record data to identify antibiotic use among hospitalized patients. *JAMA Netw. Open* **4**, e213460 (2021).
59. Talukder, A. & Ahammed, B. Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. *Nutrition* **78**, 110861 (2020).
60. Adegbosin, A. E., Stantic, B. & Sun, J. Efficacy of deep learning methods for predicting under-five mortality in 34 low-income and middle-income countries. *BMJ Open* **10**, e034524 (2020).
61. Fenta, H. M., Zewotir, T. & Muluneh, E. K. A machine learning classifier approach for identifying the determinants of under-five child undernutrition in Ethiopian administrative zones. *BMC Med. Inform. Decis. Mak.* **21**, 291 (2021).
62. Methun, M. I. H., Kabir, A., Islam, S., Hossain, M. I. & Darda, M. A. A machine learning logistic classifier approach for identifying the determinants of under-5 child morbidity in Bangladesh. *Clin. Epidemiol. Glob. Heal.* **12**, 100812 (2021).
63. Kurowski, B. G. et al. Electronic health record and patterns of care for children with cerebral palsy. *Dev. Med. Child Neurol.* **63**, 1337–1343 (2021).
64. Chung, Y. G., Jeon, Y., Yoo, S., Kim, H. & Hwang, H. Big data analysis and artificial intelligence (AI) in epilepsy – common data model analysis and machine learning-based seizure detection and forecasting. *Clin. Exp. Pediatr.* **65**, 272–282, <https://doi.org/10.3345/CEP.2021.00766> (2022).
65. Rogerson, S., Crowley, P. A., Crowley, S., Kohane, I. & Chou, J. The Harvard Necrotizing Enterocolitis Database: an artificial intelligence-friendly data repository with a unique ontology. *J. Pediatr. Gastroenterol. Nutr.* **67**, S146 (2018).
66. Colman, R. J., Dhaliwal, J. & Rosen, M. J. Predicting therapeutic response in pediatric ulcerative colitis—a journey towards precision medicine. *Front. Pediatr.* **9**, 19 (2021).
67. Patel, D. et al. Does machine learning have a role in the prediction of asthma in children? *Paediatr. Respir. Rev.* **41**, 51–60 (2022).
68. Jackson, N. et al. Machine learning analysis of airway transcriptomic data identifies distinct childhood asthma endotypes. *Am. Thorac. Soc. Int. Conf. Meet. Abstr.* A1151–A1151 (2021).
69. Oliverio, A. L. et al. Validating a computable phenotype for nephrotic syndrome in children and adults using PCORnet data. *Kidney360* **2**, 1979–1986 (2019).
70. Ranchin, B., Maucourt-Boulch, D. & Bacchetta, J. Big data and outcomes in paediatric haemodialysis: how can nephrologists use these new tools in daily practice? *Nephrol. Dial. Transplant.* **36**, 387–391 (2021).
71. McKinney, E. F., Lee, J. C., Jayne, D. R. W., Lyons, P. A. & Smith, K. G. C. T-cell exhaustion, co-stimulation and clinical outcome in autoimmunity and infection. *Nature* **523**, 612–616 (2015).
72. Zhao, M., Xu, S., Cavagnaro, M. J., Zhang, W. & Shi, J. Quantitative analysis and visualization of the interaction between intestinal microbiota and type 1 diabetes in children based on multi-databases. *Front. Pediatr.* **9**, 1389 (2021).
73. Webb-Robertson, B. J. M. et al. Prediction of the development of islet auto-antibodies through integration of environmental, genetic, and metabolic markers. *J. Diabetes* **13**, 143–153 (2021).
74. Tayefi, M. et al. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdiscip. Rev. Comput. Stat.* **13**, e1549 (2021).
75. Househ, M. & Aldosari, B. The hazards of data mining in healthcare. *Stud. Health Technol. Inform.* **238**, 80–83 (2017).
76. Kruse, C. S., Goswamy, R., Raval, Y. & Marawi, S. Challenges and opportunities of big data in health care: a systematic review. *JMIR Med. Inform.* **4**, e38, <https://doi.org/10.2196/MEDINFORM.5359> (2016).
77. Househ, M. S., Aldosari, B., Alanazi, A., Kushniruk, A. W. & Borycki, E. M. Big data, big problems: a healthcare perspective. *Stud. Health Technol. Inform.* **238**, 36–39 (2017).
78. Kayaalp, M. Patient privacy in the era of big data. *Balk. Med. J.* **35**, 8–17 (2018).
79. McCafferty, C., Chaaban, J. & Ignjatovic, V. Plasma proteomics and the paediatric patient. *Expert Rev. Proteomics* **16**, 401–411, <https://doi.org/10.1080/14789450.2019.1608186> (2019).
80. Bardanzellu, F. & Fanos, V. How could metabolomics change pediatric health? *Fanos Ital. J. Pediatr.* **46**, 37 (2020).
81. Lalonde, E. et al. Genomic diagnosis for pediatric disorders: revolution and evolution. *Front. Pediatr.* **8**, 373 (2020).

ACKNOWLEDGEMENTS

The authors acknowledge Chiranjeet, at AIIMS Kalyani.

AUTHOR CONTRIBUTIONS

K.M. and S.C.M. together conceptualized and designed the article. K.M. helped in acquisition of data and interpretation of data. S.C.M. drafted the article and K.M. revised it critically for important intellectual content. S.C.M. and K.M. approved the final version.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Saheli Chatterjee Misra.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.