16 Analytics

Suranga N. Kasthurirathne and Shaun J. Grannis

Learning Objectives
At the end of the chapter, the reader should be able to:

- Identify and define key terms and concepts associated with data analytics and big data.
- Identify and characterize primary and secondary data sources of relevance to clinical informatics.
- Understand and contrast the functionality, advantages, disadvantages, and uses of natural language processing, supervised and unsupervised learning approaches, and neural networks.
- Investigate the role and value of various visualization techniques in clinical informatics.
- Evaluate machine learning approaches using performance metrics such as precision, recall, accuracy, and Area under the ROC curve (AUC ROC).
- Identify practical considerations and implications that influence the adoption of analytical tools and methods.

Practice Domains
Domain 1: Fundamental Knowledge and Skills

- K004. Descriptive and inferential statistics.
- K025. The flow of data, information, and knowledge within the health system.

Domain 2: Improving Care Delivery and Outcomes

- K049. Prediction models.
- K050. Risk stratification and adjustment.

Domain 4: Data Governance and Data Analytics

- K101. Definitions and appropriate use of descriptive, diagnostic, predictive, and prescriptive analytics.
- K102. Analytic tools and techniques (e.g., Boolean, Bayesian, statistical/mathematical modeling).
- K103. Advanced modeling and algorithms.
- K104. Artificial intelligence.
- K105. Machine learning (e.g., neural networks, support vector machines, Bayesian networks).
- K106. Data visualization (e.g., graphical, geospatial, 3D modeling, dashboards, heat maps).
- K107. Natural language processing.

Case Vignette
You have just been appointed as the Chief Medical Information Officer (CMIO) of a large hospital system with an established medical record platform that has been used to capture patient data for several years. Your CEO has heard of the benefits of data analytics in informing healthcare delivery. She has tasked you with putting together a long-term plan for adopting analytics into your health system. How would you approach this challenge?

This chapter was adapted from a prior publication [1].

Introduction

Data plays a significant role in modern society and the economy. As envisioned by mathematician Clive Humby, credited with the phrase ‘data is the new oil’ [2], it continues to be of unparalleled value in driving the information age. Increased uptake of health information systems has led to increased accessibility and availability of health-related datasets. However, learning how to leverage various complex heterogeneous datasets to infer value in clinical settings is an uphill task. For the last several decades, researchers have demon-

S. N. Kasthurirathne (✉) · S. J. Grannis
Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, USA

Indiana University School of Medicine, Indianapolis, USA
e-mail: snkasthu@iu.edu

J. T. Finnell, B. E. Dixon (eds.), *Clinical Informatics Study Guide*, https://doi.org/10.1007/978-3-030-93765-2_16

strated the ability to apply various analytical methods in response to multiple challenges impacting various clinical care domains [3–6]. Adoption of such analytical tools at scale is hampered by concerns of algorithmic bias and unfairness [7], limited generalizability and transportability of models across new patient populations and settings, and challenges in implementation and quality control [8–10].

However, these barriers are subsiding. Widespread acceptance of analytical methods and increasing demands on the clinical workforce have paved the way for ramping up efforts to develop and deploy innovative analytical solutions to address a range of use cases, including data analysis, machine learning, risk assessment and stratification, and visualization.

However, keeping up with the rapidly evolving Artificial Intelligence domain and apply these concepts to clinical informatics. Further, understanding the plethora of primary and secondary data sources captured at the patient- and population-level and leveraging these datasets to extract and model clinical, behavioral, and social determinants influencing patient health and wellbeing can be challenging. Thus, researchers and practitioners of clinical informatics need to obtain a firm grounding in the fundamentals of analytics and potential limitations and challenges that must be overcome to build robust analytical solutions. This chapter provides a detailed overview of theoretical and practical aspects of data analytics and its application in clinical informatics.

Data to Wisdom

This section provides a brief overview of data, information, knowledge, and wisdom and their relationships. It further offers an introduction to other analytical terms and the categorization of various analytical methods.

To learn about data analytics and the use of big data, readers must first understand several key terms - data, information, knowledge, and wisdom - and the hierarchical relationships between them. These relationships are described as the *DIKW* (Data, Information, Knowledge, Wisdom) pyramid (Fig. 16.1a).

- *Data* (base of the pyramid) collects discrete, objective facts or observations that are represented in raw and unorganized form without context. As such, they are of little value. As an example, the string ‘101’ is discrete and objective but lacks any context.
- *Information* is organized or structured data that has been prepared so that it is relevant for a specific need and is therefore valid, relevant, and valuable. For example, knowing that the string ‘101’ mentioned earlier represents an adult’s body temperature in Fahrenheit adds more context and is more meaningful.
- *Knowledge* is a flux of framed experiences, contextual information, values, expert insight, and grounded intuition that offer an environment and framework for evaluating and incorporating new information [12]. For example, additional context around adult human body temperature helps a clinician understand that this patient suffers from a fever.
- *Wisdom* (the topmost point of the pyramid) is the ability to increase effectiveness. It adds value, which requires the use of judgment. Given that judgment may be influenced by an individual’s aesthetic values or ethics, wisdom is often personal and inherent. For example, by evaluating the knowledge presented previously, a clinician understands that they must address the patient’s fever.

The relationships between these factors are represented in Fig. 16.1(a) as layers of a pyramid, with the largest source (data) at the very bottom and the smallest source (wisdom) at the very top. Analytics (defined below) help researchers advance from data (widely available but low value) to information, knowledge, and wisdom (increasingly harder to obtain and more valued).

Key Terms in Analytics

Definitions for several key terms used in the analytics domain are as follows:

- *Data science*: The multi-disciplinary field leverages various methods, processes, and algorithms to extract knowledge and insights from structured and unstructured data.
- *Artificial intelligence (AI):* A subdomain of computer science that focuses on the simulation of human intelligence (or brain function) by a machine. AI is a broad domain encompassing machine learning and other topics, such as logic, problem-solving, and reasoning, which are out of scope for this chapter.
- *Machine learning:* The ability of a computer system to learn from the external environment or a data source to improve its ability to perform a task. These approaches enable various algorithms to learn from data without any explicit programming. Machine learning is a subset of AI.
- *Analytics*: The discovery, interpretation, and communication of meaningful patterns found in data, as well as the application of data patterns for effective decision-making.

Descriptive and Inferential Statistics

Descriptive Statistics Descriptive statistics refer to a group of analytical methods used to summarize datasets in a manner that ‘describes’ or summarizes a population, making them easily interpretable to researchers. Descriptive statistics are calculated using basic mathematics and statistics measures such as percentages, mean, median, and mode val-

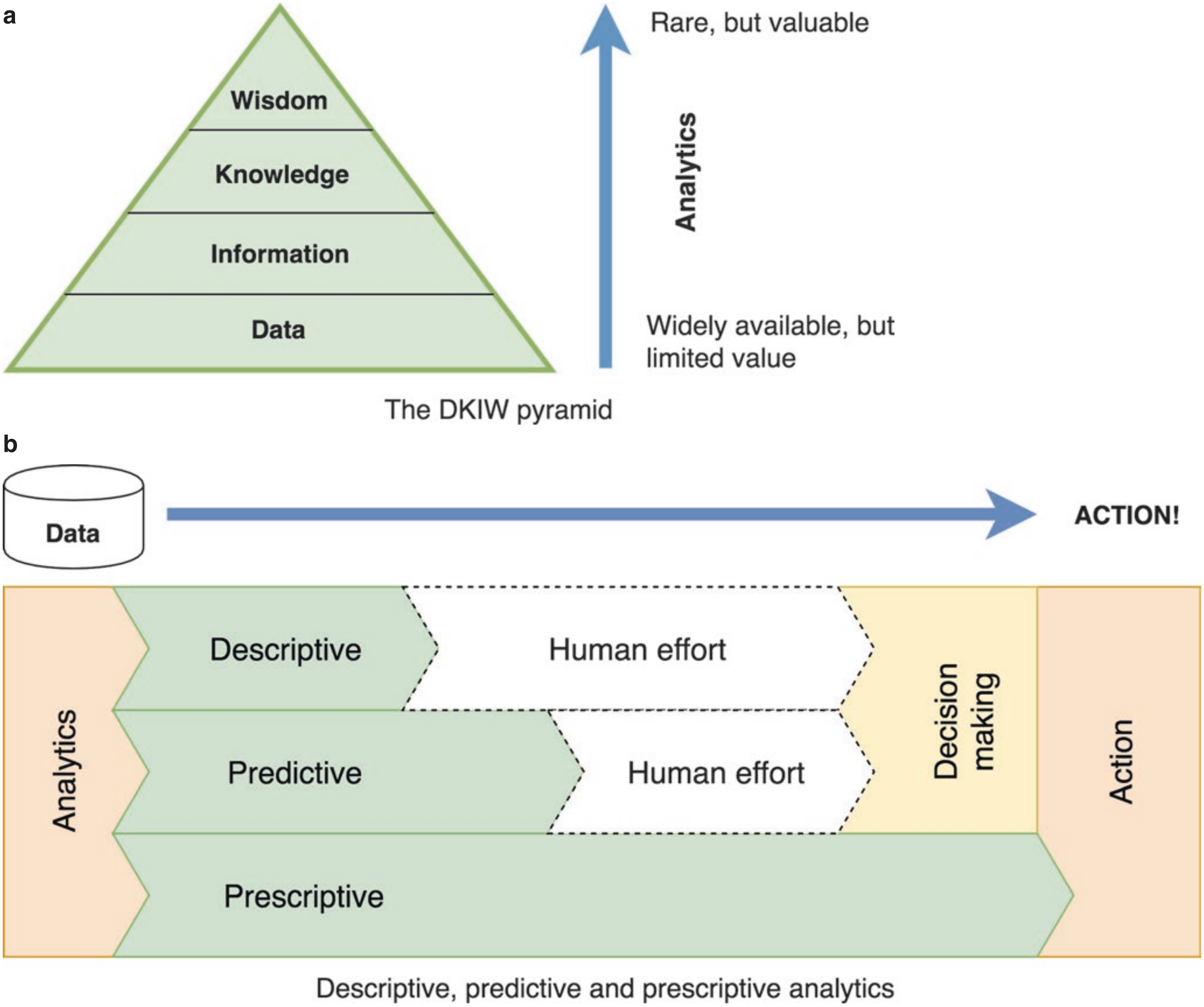

Fig. 16.1 Introduction to fundamental concepts: (**a**) the DKIW pyramid and (**b**) descriptive, predictive, and prescriptive analytics. (**b**) is derived from Gartner Inc's analytical capabilities visualization [1, 11]

ues [13]. These metrics do not allow us to make conclusions beyond the datasets under test or reach conclusions regarding any hypothesis.

Inferential Statistics In contrast to descriptive statistics, Inferential statistics reach conclusions beyond the dataset under test [14]. Inferential statistics uses a random sample of data extracted from a broader population to describe and infer a larger population of interest. As such, inferential statistical methods are essential where assessing all individuals in a patient population is infeasible. In such an event, inferential methods can make generalizations across the broader population of interest. Standard inferential statistical methods include hypothesis tests, confidence intervals, and regression analysis.

Data Sources

Data sources for clinical informatics research can be categorized into *primary* and *secondary* sources. Primary data sources are datasets collected by healthcare providers for the specific purpose of providing healthcare. Most primary data collection activities are performed using Electronic Health Record systems, lab information systems, or other medical testing equipment. Secondary data sources include existing data that were collected for other purposes. There are three main types of secondary data sources for clinical informatics research:

1. **Surveys**. A valid, commonly used method to collect demographic information, personal behaviors, and attitudes. In some cases, data from physical examinations and laboratory tests are collected in addition to these self-

reported data. Surveys are typically utilized to collect data for questions that cannot be answered from other data sources. They can be primary or secondary data sources depending on who collected the data. Generally, surveys are conducted at national (population-based), state, and local levels. These data are collected at a personal or population level, depending on the sampling methodology. Widely used secondary survey data sources include the National Health and Nutrition Examination Survey (NHANES) [15] and the Behavioral Risk Factor Surveillance System (BRFSS) [16].

2. **Registries**. A method to collect data and information on individuals suffering from specific diseases or conditions. The Agency for Healthcare Research and Quality (AHRQ) defines four types of registries [17]: (1) product (i.e., pharmaceutical, medical devices, or diagnostic/therapeutic equipment), (2) health services (i.e., exposure to medical procedures, clinical encounters, or hospitalizations), (3) disease or condition (i.e., all patients have same disease or condition), and (4) a combination of any or all of the above. Data stored in these registries are collected in a standardized, predefined method specific for each registry.
3. **Health services data**. These data are collected as part of the routine healthcare processes. They contain large samples of individualized patient data, including diagnoses, medications, procedures, imaging, and medical notes. The two main types of health services data are administrative claims data and data extracted from medical records. Administrative claims data are data which encode for diagnoses, medications, and procedures for billing purposes. They are coded using standard coding systems such as the International Classification of Diseases (ICD) [18], Systematized Nomenclature of Medicine (SNOMED®) [19], Current Procedural Terminology (CPT) [20], and Logical Observation Identifiers Names and Codes (LOINC®) [21]. However, most data collected during the healthcare process are *unstructured* (i.e., not coded) and exist as free-text, images, or video. Thus, manual medical chart reviews through Electronic Health Records (EHRs) are needed to extract and codify the data before analyses. In the past several years, developments in Natural Language Processing (NLP) and machine learning have shown promise in extracting value from unstructured clinical data. Despite the potential of NLP and machine learning, widespread adoption is limited due to concerns about their overall accuracy, lack of trust by clinicians, and generalizability. There are significant limitations to consider before utilizing health services data.
4. **Big data**. Big data refers to the field of research, methodology, and expertise on the extraction, analysis, and persistence of datasets that are too large and/or complex to be analyzed by traditional methods. Thus, what constitutes big data may be specific to an individual, implementation, or location. Three concepts drive the definitions and assessment of big data, often referred to as the three V's:
   - *Volume* (quantity of data),
   - *Variety* (types of datasets), and
   - *Velocity* (how often the data are being captured/reported) [22].

Big data for clinical care delivery became feasible due to

(a) increasing adoption of Health Information System (HIS) infrastructure, which enabled widespread collection and management of patient-level data,
(b) increased interest in the collection and dissemination of population-level datasets and Geographical Information Systems (GIS) that describe a wide variety of socioeconomic measures, and
(c) reduced technical barriers and costs associated with data persistence and management.

The advent of big data brings new challenges in translating datasets of various quality, quantity, and velocity into actionable information, and ultimately, to knowledge. Big data analytics seeks to leverage improvements in computer science to address these needs. Big data are of significant interest to the clinical informatics domain due to their ability to provide broad insight into various patient health and wellbeing perspectives.

Data Pre-Processing: What Pre-Processing Steps Are Necessary to Convert a Data Set into a Format Suitable for Analytics?

Any analytical process is only as good as the quality of datasets used. As such, it is essential to ensure that datasets used for analysis are cleaned and parsed to present a concise, valid, and clear picture of the clinical scenarios or patient populations under test. Often, raw data must be transformed into **data vectors**, which can be defined as collections or arrays of numbers structured in a manner that helps an analytical approach identify relationships and patterns within the data.

Introduction to NLP

Natural language processing (NLP) is a domain of AI which focuses on how computers interpret written and spoken human language. NLP is particularly relevant to clinical care initiatives, given that clinical reports consist of up to 80%

Table 16.1 Basic NLP Techniques

Technique	Description	Example
Lemmatization	Grouping together inflected forms of a word so they may be analyzed as a single item.	'Good' is the lemmatized form of 'better.'
Stemming	The task of reducing inflected or derived words into their root form to better identify various uses of a single word.	'Walk' is the stem of 'walking.'
Summarization	Produces a readable summary of a larger block of text.	Condensation of a paragraph or a larger text document into a smaller set of sentences or shorter paragraphs.
Sentence boundary detection	The task of identifying the boundaries (start and end) of sentences from within a larger text document.	Identify that the use of a dot in the term 'Dr. Walker' is intended to abbreviate the term 'doctor' and not intended as a sentence break.
Sentiment analysis or opinion mining	Identify affective states and subjective information used to infer polarity on specific topics.	Identify that the phrase 'negative for cancer' indicates that the patient does not have cancer.
Part-of-speech (POS) tagging	For each word within a sentence, determine the part of speech, such as verbs, nouns, or adjectives.	Identify all nouns and pronouns in a sentence.
Named entity recognition (NER), aka entity identification	The task of locating and classifying named entities into various predefined categories	Identify that 'car' is a type of 'vehicle' and that 'syphilis' is a type of 'illness'.

unstructured data [23]. NLP methods may also be used to augment the quantity and quality of structured datasets. Early attempts at NLP centered on regular expression (regex) based matching, hard-coded rule systems, and decision trees, rigidly tied to specific use cases. As illustrated in Table 16.1, NLP methodology has expanded to support a broader range of techniques.

These NLP techniques should be applied with due consideration to the context they are applied to. For example, noting that a social worker's note refers to a historical (not current) case of homelessness or that evidence of alcohol abuse is linked to the family member of a patient rather than the actual patient plays a significant role in efficient data extraction. While tools and techniques to evaluate the context are available [24], assessing these constraints adds additional complexity to NLP in the clinical domain.

Machine Learning Approaches

The rapid advancement of Artificial intelligence, computer science, and decreasing cost of computational and storage resources has brought about significant awareness of the role of AI-driven methods for decision-making. Increasing quantities of data and the complex nature of decision-making impedes a human expert's ability to make rational decisions based on the datasets at hand. Machine learning approaches enable users to learn from these datasets more efficiently and effectively and apply this knowledge for decision-making. Machine learning efforts can be broadly categorized into *supervised* and *unsupervised* learning approaches.

Supervised Learning Approaches

Supervised learning is an approach where, given input features and an outcome variable of interest, an algorithm can learn the mapping function to convert the input features into the outcome of interest. This is referred to as supervised learning because the outcome variable serves as a gold standard used to guide the algorithm on a learning process. However, using these methods can be costly and resource-intensive as they may require human expert input for defining and preparing a gold standard. Supervised approaches can be grouped into two major categories as described below.

(a) **Classification models** Algorithms that predict a discrete or categorical output variable. Listed below are some widely known classification algorithms:
- **Simple Logistic (SL)** Given a set of training samples with a labeled outcome, SL models develop a logistic function to predict the outcome variable. SL does not rely on assumptions of normality for predictor variables. These models are very simplistic, mainly when few or no interaction terms are used [25].
- **Support Vector Machines (SVM)** Given a set of training examples with labeled outcomes, SVM identifies an optimal hyperplane (a subspace whose dimension is −1 of its ambient space) capable of separating data into each outcome. SVM models work well on small, clean datasets given the ease of drawing clear hyperplanes across these datasets. However, they are less effective on larger, noisier datasets with multiple overlapping classes.
- **Bayesian classifiers** are probabilistic classifiers based on Bayes' theorem, which describes the probability of an event based on prior knowledge of conditions related to the event. This approach assumes that all features in a model are independent and that the presence of one feature does not impact the presence

of another. Given this assumption, their use may be somewhat restricted in the healthcare domain.

- **Decision trees** A supervised learning approach seeks to predict an outcome's value by learning decision rules inferred from the training dataset. Decision trees are simple to interpret and require little data preparation and cleaning. They can also be used for both classification and regression. However, decision tree models may result in overly complex trees that are too tightly linked to training data and do not yield satisfactory performance across other datasets.

(b) **Regression models** Algorithms that predict a numerical continuous output variable. Examples include Simple Logistic Regression and Random Forest Regression. These algorithms mimic their peers in the classification section but are designed to output a continuous variable rather than a categorical value.

In addition, several supervised learning algorithms may be integrated to develop *ensemble models*. As discussed earlier, each supervised learning algorithm poses unique strengths and weaknesses. An ensemble model combines multiple machine learning algorithms into a single predictive model, thereby combining the advantages of each unique model towards the final outcome prediction. Decision trees are traditionally implemented as ensembles consisting of 'forests' of multiple trees. Two of the most widely known ensemble-based implementations are Random Forest [26] and eXtreme Gradient Boosting (XGBoost) [27]. Random Forest builds all trees and then averages the predictions made by each tree, while gradient boosting methods build new trees focused on addressing errors in prior trees.

Unsupervised Learning Approaches

Unsupervised (or clustering) learning approaches are methods where an algorithm learns to model the underlying distribution of data elements given input features but no outcome variable. Such approaches are data-driven and rely purely on the quantity and quality of data used in the training process. Unsupervised methods are relatively easier to train because they do not require the manual cost and effort needed to develop a gold standard. However, this usually leads to weaker performance. Listed below are two widely used clustering algorithms.

- **k-means clustering** An approach that seeks to group each observation into a subset of clusters where each observation belongs to the cluster with the nearest mean value. k-means are one of the oldest and widely used clustering algorithms. They are efficient and straightforward and therefore suitable for large-scale datasets. However, the algorithm cannot pre-determine an optimal number for k, meaning that the best value must be selected via incremental evaluation using multiple k values.
- **Hierarchical clustering** An approach that seeks to build out a hierarchy of clusters. They can be agglomerative (each individual instance starts as a separate cluster, with pairs of clusters merging as instances traverse up the hierarchy) and divisive (all observations start with one cluster, and splits are performed as instances traverse down the hierarchy). While descriptive, this approach is more complex and requires more memory. Thus, it may be unsuitable for larger datasets.

Neural Networks

Neural networks are computing systems inspired by the biological neural networks that constitute animal brains. A neural network consists of layers of connected nodes that are referred to as neurons. Neural networks can be either supervised or unsupervised by nature. Although the principles of neural networks were known for decades, they did not achieve mainstream interest and adoption until large quantities of data and computational processing resources that unleashed their true potential became readily available. At a minimum, a neural network consists of three layers; one input layer, a hidden layer (any layer located between the input and output layers), and an output layer. A neural network with more than a single hidden layer is referred to as a deep learning network. In contrast to other classification systems, neural networks outperform traditional machine learning approaches as the scale of data increases.

Various neural network systems have been implemented in response to a myriad of challenges. However, they are increasingly complex, making them harder to interpret than other classification models. This limits the application of neural networks in certain healthcare domains where the interpretability of a prediction is of significant importance. However, they are invaluable in analyzing images, data streams, and genomic datasets. Currently, neural networks are widely used for various tasks linked to clinical care delivery [28–30]. Listed below are several commonly used classes of neural networks.

- **Convolutional Neural Network (CNN)** A neural network approach focused on challenges involving visual imagery. These are commonly applied to analyzing images or videos.
- **Recurrent Neural Network (RNN)** Neural network approach where connections between nodes form a

Table 16.2 Potential clinical informatics-related use cases and hypothetical analytical solutions

Use case	Potential solution
Predict patient-level hospital readmission rates	Patient-level hospital readmissions may be predicted using existing clinical, behavioral, and demographic datasets and supervised learning [29] or neural network-based methods.
Identifying types of patients most likely to develop opioid addictions.	To effectively address the causes of opioid addiction, it may be useful to identify different subpopulations of patients at most risk. A variety of basic descriptive statistics, risk stratification methods, or more complex predictive modeling approaches may be used for this purpose.
Identification of notifiable conditions for public health reporting	Free text reports may be searched for evidence of notifiable conditions using basic string search functions, regular expressions, or other more complex NLP-driven methods.
Support clinicians in cancer detection	Deep learning models can be trained to detect a variety of cancers using patient CT scans.
Detect drug-drug interactions	A variety of methods ranging from basic rule-based systems to deep learning models and neural networks can be applied to identify potentially harmful drug-drug interactions

directed graph representing temporal sequences, allowing the model to exhibit dynamic temporal behavior. These are commonly used to predict sequences of events such as changing stock prices, patient heart rate, or other sequential measures.

- **Long short-term memory (LTSM)** A form of RNN capable of learning long-term dependencies, thereby enabling it to support sequential predictions. These are applied to similar use cases as RNN's.

Applications of Analytics in Clinical Informatics

In this section, we briefly discuss several use cases where analytics could be applied to clinical informatics. Table 16.2 presents several clinical informatics-oriented use cases that can be addressed using analytics.

Common Pitfalls and Challenges

Some common pitfalls and challenges associated with machine learning are as follows:

- **Overfitting** A decision model is said to overfit if it captures the underlying structure of a dataset too stringently and thus, fails to achieve consistent performance across a different dataset. Overfitting is caused by noise (irrelevant or incorrect data elements included in the dataset) that does not generalize across other datasets. Ensuring that a model does not suffer from overfitting is referred to as *generalization*. To protect against overfitting, models should be trained using broad representative datasets and evaluated across various heterogeneous patient populations.
- **Underfitting** A decision model is said to be underfitting if it cannot adequately capture the underlying structure of a dataset and thus, underperforms against both the current and other datasets. To protect against underfitting, models should be trained using various features that adequately represent a use case under test. Often, researchers must deal with a tradeoff between model overfitting and underfitting in delivering effective models.
- **Class imbalance** In many real-world classification problems, the outcome variable (class) may not make up an equal or reasonable proportion of the dataset. For example, the prevalence of HIV, AIDS, or Rabies may be significantly low across the general population. A classification model may not have enough 'signal' in the training data to deliver adequate predictive performance in such a scenario. Two sampling methods may be considered to address class imbalance; o*versampling* (supplementing the minority class/es with copies of minority instances) and *undersampling* (removing instances of the majority class at random to improve class balance).

Model Training, Evaluation, and Validation

Researchers may select from multiple model training approaches based on the availability of data. Below are three training methods; train and test, cross-validation and train, validation, and test (Fig. 16.2).

Model Training Approaches

- **Train and test method** A dataset is randomly split into two sets, a larger training dataset used to train a decision model and a smaller test dataset used to test the newly trained model. Based on dataset size and quality, a training dataset could range from 70–90% of the original dataset.
- **Cross-validation** A resampling approach where the dataset is split into k many randomly selected subsets, where the user defines the size (k). We randomly choose one of the k subsets as the test dataset while the remaining subsets as training datasets. A model is trained using the (k–1) training datasets and evaluated using the test datasets. This process is carried out k-many times, and performance

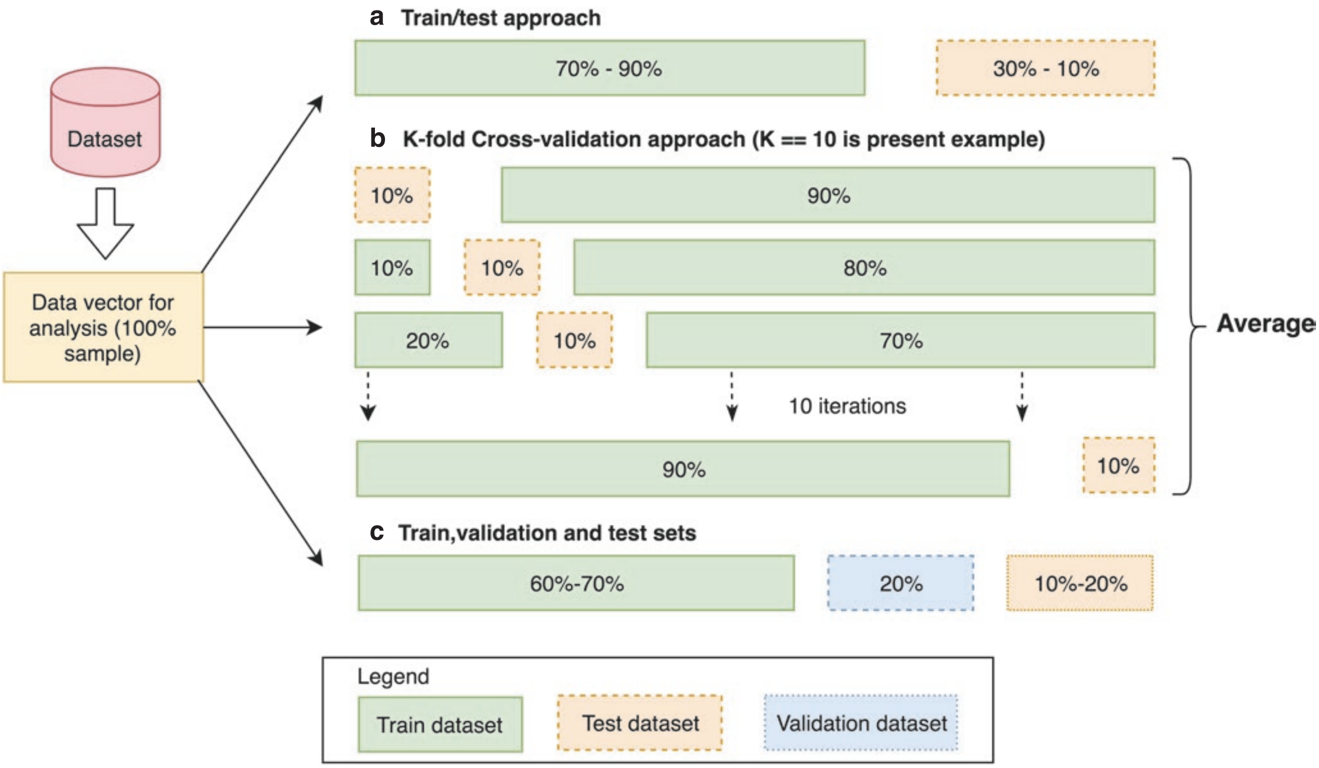

Fig. 16.2 Comparison of various training, validation, and testing methods for maching learning. (**a**) the dataset is divided into training and test subsets. (**b**) the dataset is resampled multiple times using random subsets of data. (**c**) the dataset is divided into training, validation, and testing subsets [1]

results for each iteration are averaged to produce less variable performance results. k = 10 is widely used, but values as small as five may be used based on the dataset at hand. Cross-validation methods are traditionally used on smaller datasets that require optimal use of data for training. However, this approach is vulnerable to overfitting.

- **Train, validation, and test sets.** This newer approach is more suitable for situations where a significant quantity of data is available. The dataset is randomly split into train, validation, and test sets. The training dataset is used to train the decision model. The validation dataset is then used to iteratively test the decision model and update its parameters for optimal performance. Once model parameters have been configured for optimal results, the model is evaluated using the holdout test dataset.

Performance Metrics

It is essential to evaluate the performance of a decision model using a variety of performance metrics.

- *Sensitivity (AKA recall)*: Proportion of actual positives that are correctly identified.
- *Specificity (AKA true negative rate):* Proportion of actual negatives that are correctly identified.
- *Precision (AKA positive predictive value):* Proportion of positive identifications that are correct.
- *F1-score*: Accuracy measure representing the *harmonic mean* (an average used for numbers that represent rate or ratio) between precision and recall
- *The Area Under the Receiver Operator Characteristic curve (AUC ROC)*: The Receiver Operator Characteristic (ROC) is a graphical plot that demonstrates the diagnostic performance of a classification model across various threshold configurations. The AUC ROC score measures the two-dimensional space underneath the ROC curve. Thus, the AUC ROC score can range between 0 (minimum) and 1 (maximum). An AUC ROC of 0.5 indicates that a model has no discrimination power.

It is also essential to identify the most appropriate performance metrics to evaluate model performance given a specific use case. For example, analytical methods that seek to predict the probability of high risk or high-cost events such as mortality or permanent injury should optimize sensitivity to increase the chances of identifying as many patients in need as possible. In contrast, solutions to identify the likelihood of a less risky event may focus on optimizing precision to reduce the burden of false positives on clinicians. Often it is helpful to compare tradeoffs between different perfor-

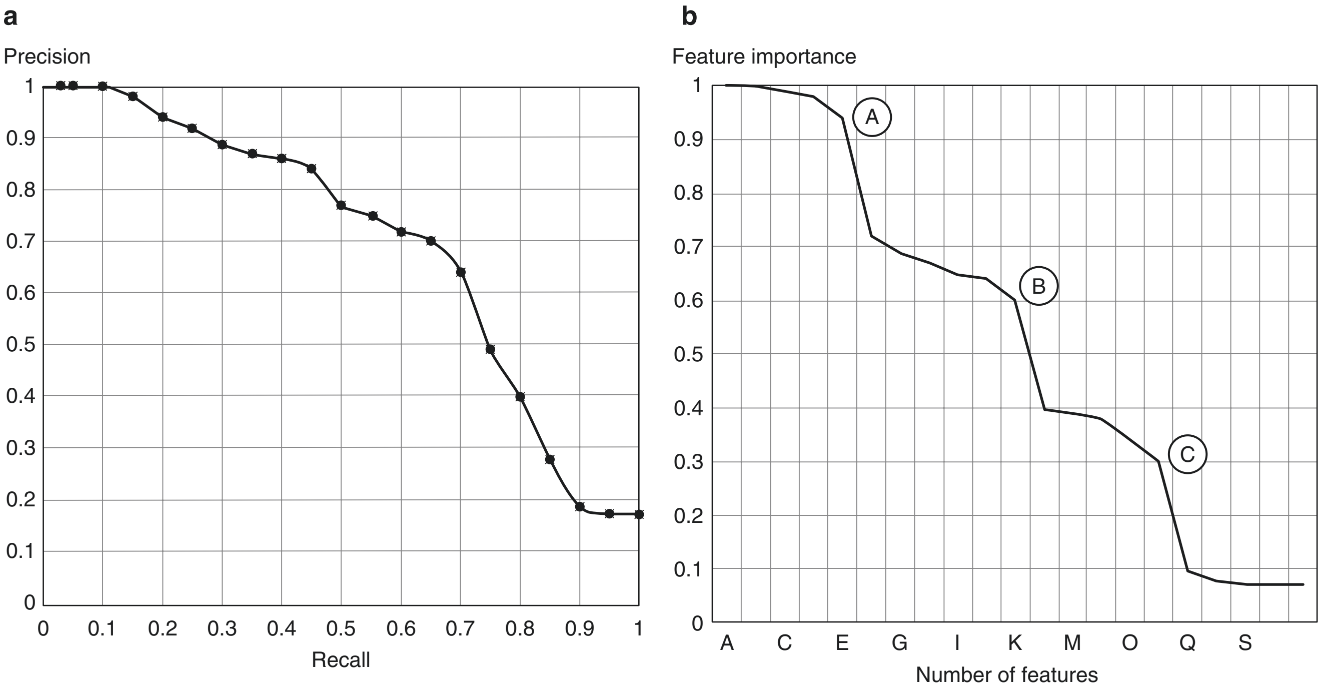

Fig. 16.3 (**a**) Precision-Recall curve demonstrating tradeoff between precision and recall at different cutoff thresholds, and (**b**) feature importance scores for each feature (ranked from the most important to least important). Adapted from [1]

mance measures to select the optimal model. Precision-recall curves compare variations in each metric across different cutoff thresholds, thereby enabling researchers to identify optimal thresholds based on the predictive performance of their choice. The sample precision-recall curve in Fig. 16.3 (**a**) presents variations of precision against recall. For example, this plot informs a researcher that 0.65 is the maximum precision achievable for a recall of > = 0.7. However, the maximum precision for a recall of > = 0.3 is 0.9.

Feature Selection Techniques

Feature selection, alternatively known as attribute selection or variable selection, is the process of selecting the most relevant features with the potential to contribute most towards a machine learning task. Proper feature selection can lead to numerous benefits, including reduced risk of overfitting, improved model accuracy, and reduced model training time and hardware requirements. Primary feature selection can be performed via manual review. A human expert with knowledge on a particular topic manually reviews a list of features and selects a subset of potentially relevant features based on their expertise. However, manual review becomes challenging with larger feature sets and also when investigating a lesser-known domain. Automated feature selection methods can be applied to address these situations. Automated feature selection methods can be classified as filter, wrapper, and embedded methods.

- *Filter methods*: Approaches that apply statistical measures to assign a score to each feature.
  - *Univariate selection*: Selects features with the strongest relationships to the outcome variable.
  - Information gain (AKA Kullback-Leibler divergence) [31]: Evaluates a feature's worth by measuring the information gain to the outcome of interest.
- *Wrapper methods*: Approaches that consider feature selection as a search problem using various combinations of features. Recursive Feature Elimination (RFE) is a greedy optimization (an approach that seeks to make a locally optimal choice at each stage) to identify the feature set with the best model performance by iteratively creating models.
- *Embedded methods*: Approaches that identify which features contribute most to the model's accuracy during its training process. Learning algorithms that support embedded feature selection perform feature selection as part of the model development process. Regularization or penalization methods such as the least absolute shrinkage and selection operator (LASSO), Elastic Net, and Ridge Regression commonly use embedded feature selection methods.

An example of automated feature selection would be when clinical data elements collected from an Electronic Health Record system or Health Information Exchange are being used to perform syndromic surveillance. In such an event, these filter methods can be applied to identify a smaller subset of the most relevant features to be used in machine learning, thereby rendering the models more simplistic, with a lesser risk of overfitting. Further, a model that requires a limited number of features would be easier to operationalize in a clinical setting.

Figure 16.3(b) plots the importance of each feature, as identified by the feature selection method being used. How would a researcher identify the optimal number of features for model development? This depends on the ability to reach suitable performance metrics and considerations on model complexity. Assume that a model trained using the top 25 features (point A in Fig. 16.3(b)) does not yield adequate performance. In that case, expanding to include the top 55 features (point B in the plot) may do so. Alternatively, expanding further to include the top 85 features (point C in the plot) may be necessary. However, the inclusion of additional features results in a more complex model and may risk overfitting. Does the performance increase achieved by expanding to include extra features justify the risk of overfitting and increased complexity for the use case under study?

Model Validation

The core purpose of developing analytical models is to leverage them to predict outcomes for unseen populations. To do so, a model must demonstrate reasonable validity. A model is said to be valid if it demonstrates both internal validity and external validity.

- *Internal validation*: Testing a model's ability to replicate its predictions across the same population used to train the model. Internal validation can be performed by applying a model to a holdout dataset extracted from the original population and evaluating it using the performance metrics listed previously.
- *External validation*: Evaluate model performance against a dataset sampled from an alternative population not used in the initial training process. External validation serves as the gold standard for evaluating decision model performance. Unfortunately, external validation methods are rarely used due to the lack of access to datasets and the cost of performing such validation in the clinical domain.

Risk Stratification and Adjustment

Risk stratification is where clinicians assign patients to different tiers based on factors contributing to adverse health outcomes [31]. Different stratification methods may be selected based on each use case; as an example, one method might prioritize patients in most need (are sickest), while another may prioritize patients who are most likely to improve with care. Stratification methods result in distinct groups of patients with similar complexity and care needs. They help providers identify and mitigate patients' risks, effectively allocate healthcare delivery resources, and prioritize care for the right patients. In the most basic terms, risk stratification can be performed by using descriptive statistics to identify levels of risks and care needs based on patient demographics and the presence of chronic conditions. However, more successful approaches to risk stratification rely on complex predictive analytics [32] and phenotyping methods [33]. Risk stratification approaches are critical given value-based healthcare, which seeks to improve care outcomes while eliminating inefficiencies and reducing costs.

Data Visualization

Health systems process and analyze vast quantities of diverse datasets at a rapid pace. Effective mechanisms are needed to communicate the results of such analysis in a concise, easily understandable manner. **Data visualization** is an interdisciplinary field that integrates statistical and computing skills with design skills to enable the graphic representation of data and information. It helps reduce the burden of decision-making using complex datasets.

Basic types of health data visualization methods include various types of charts, tables, maps, scatter plots, timelines, and infographics created using various office application packages such as Microsoft Office or Apache Open Office. Alternatively, 3-Dimensional (3D) visualization techniques are widely used in clinical informatics to offer a clear rendering of the functionality of complex organs such as the human heart and aid in the diagnosis and effective delivery of various oncology, cardiology, and neurology procedures. Alternatively, geospatial visualization techniques can integrate relevant clinical or health information to geographic locations such as latitude and longitude, census tract, zip code, county, state, or country [34].

Data dashboards that incorporate one or many of these methods are used to visualize more complex datasets and interpretations in an easily accessible manner. Such dashboards may represent operational data (operational dashboards), presenting a real-time assessment of the use case under test, or strategic dashboards, representing trends or changes over time. Powerful, specialized tools such as Tableau or Power BI are widely used to create interactive dashboards or may be updated in real-time or at regular intervals. Notably, a variety of such dashboards were developed in response to the COVID-19 pandemic. Examples include dashboards built atop Indiana's statewide health information exchange for population-level surveillance and in support of pandemic response efforts across communities [35], as well as dashboards developed by Johns Hopkins University to provide timely information on COVID-19 cases and deaths worldwide [36].

Emerging Trends

The future of AI, particularly in the healthcare domain, continues to evolve in response to significant technological advances, uptake of tools and information systems, and emerging awareness of its value in driving healthcare delivery and outcomes. We highlight several notable trends in the clinical analytics domain.

Democratizing Access to Datasets for Effective Analytical Efforts

Widespread adoption of HIS has resulted in increased efforts to collect and curate clinical data. However, regulatory frameworks enforced by many countries limit the sharing of protected health information outside healthcare organizations. Limited or burdensome data access hinders the reproduction, sharing, and re-use of machine learning solutions across larger audiences and restricts inter-organizational collaboration addressing various healthcare challenges and building generalized machine learning models targeting diverse populations.

Efforts to enable better access to data include creating standardized data lakes with tools for effective access, use, and analytical efforts. Such attempts include the Observational Health Data Sciences and Informatics (OHDSI) initiative [37], a multi-stakeholder collaborative which seeks to improve health by empowering a community to collaboratively generate evidence that promotes better health decisions and better care, and currently boasts access to 600 million patients spread across 30 countries [38], as well as the National COVID Cohort Collaborative (N3C), which seeks to bring together clinical data and expertise from across the US to answer critical research questions to address the COVID-19 pandemic [39]. Other efforts involve using advanced analytical approaches such as Generative Adversarial Network (GAN) models to create large, realistic synthetic datasets that mimic original data sources but offer limited risk of re-identification [40, 41].

Awareness of Biases Present in Analytical Models

Most datasets used in healthcare research are not originally collected for research purposes [42, 43]. Such datasets are susceptible to **biases**, defined as systematic errors caused by prejudiced decision-making, poor representation of vulnerable populations, and incomplete data collection errors [44, 45]. Biases place privileged groups at a systematic advantage over unprivileged groups [44].

If used for analytics, such datasets may lead to the garbage in - garbage out problem [46], resulting in biased models harmful to vulnerable populations such as racial and ethnic minorities, older adults, or persons with special healthcare needs [47–49]. Biases can also be harmful to individuals with negative Social Determinants of Health (SDoH), defined as conditions in which people are born, grow, live, work, and age [50]. Such biases may present significant harm to patients and result in legal penalties and negative attention to healthcare systems [44]. There is increased awareness of the need to effectively identify and mitigate biases present in analytical models via effective data collection and curation methods that improve data quality and other analytical methods that improve fairness in models trained using messy data.

Summary

The popularization and adoption of analytical approaches for the healthcare domain continues at a rapid pace. To keep up with these advances, clinical informaticians must obtain a firm grounding in the fundamentals of analytics and the potential limitations and challenges that must be overcome to build and maintain robust analytical solutions. This chapter provided (a) a detailed description of the nature of data, information, wisdom, and knowledge, (b) key definitions associated with data analytics, (c) introduction to various machine learning algorithms, their advantages, and limitations, and (d) various evaluation methods to assess analytical performance. To support clinical informaticians in leveraging these lessons for practical use, it also included content on practical considerations, limitations, and challenges that may impede the implementation of AI tools in support of clinical care delivery. These lessons serve as steppingstones for researchers who wish to become familiar with the current analytics domain and support self-learning to keep up with the latest advances.

Questions for Discussion

1. Contrast various predictive performance metrics and identify clinical scenarios where you may favor one over the others. How would you explain these choices to your clinical team?
2. Contrast neural networks, classification algorithms, and clustering algorithms. In which use cases would you prefer each of these methods over the others? Why?
3. Identify common pitfalls and challenges of applying data science and analytics in clinical practice. How are emerging trends in clinical analytics addressing or bypassing these limitations?

References

1. Kasthurirathne SN, Ho YA, Dixon BE. Public health analytics and big data. In Public Health Informatics and Information Systems. 2020. (pp. 203–219). Springer, Cham.
2. Van der Aalst WM. Data scientist: the engineer of the future. Enterprise interoperability VI. Cham: Springer; 2014. p. 13–26.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
4. Nam JG, Park S, Hwang EJ, Lee JH, Jin K-N, Lim KY, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology. 2019;290(1):218–28.
5. Ding Y, Sohn JH, Kawczynski MG, Trivedi H, Harnish R, Jenkins NW, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. Radiology. 2019;290(2):456–64.
6. Kasthurirathne SN, Vest JR, Menachemi N, Halverson PK, Grannis SJ. Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services. J Am Med Inform Assoc. 2018;25(1):47–53.
7. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53.
8. Rolka H, Walker DW, English R, Katzoff MJ, Scogin G, Neuhaus E. Analytical challenges for emerging public health surveillance. CDC's vision for Public Health Surveillance in the 21st Century. MMWR Suppl. 2012;61:35.
9. Nambiar R, Bhardwaj R, Sethi A, Vargheese R, editors. A look at challenges and opportunities of big data analytics in healthcare. In: 2013 IEEE international conference on Big Data. Piscataway: IEEE; 2013.
10. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):1–9.
11. Gartner Inc. Gartner says advanced analytics is a top business priority 2014. Available from: https://www.gartner.com/en/newsroom/press-releases/2014-10-21-gartner-says-advanced-analytics-is-a-top-business-priority.
12. Davenport TH, Prusak L. Working knowledge: how organizations manage what they know. Boston: Harvard Business Press; 1998.
13. Fisher MJ, Marshall AP. Understanding descriptive statistics. Aust Crit Care. 2009;22(2):93–7.
14. Byrne G. A statistical primer: understanding descriptive and inferential statistics. Evid Based Libr Inf Pract. 2007;2(1):32–47.
15. Centers for disease control and prevention. About the national health and nutrition examination survey 2017. Available from: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm.
16. Centers for disease control and prevention. About BRFSS 2014. Available from: https://www.cdc.gov/brfss/about/index.htm.
17. Gliklich RE, Dreyer NA, Leavy MB, Quality/AHRQ AHR. Registries for evaluating patient outcomes: a user's guide. Washington, DC: U.S. Department of Health and Human Services; 2014.
18. World Health Organization. International statistical classification of diseases and related health problems. Geneva: World Health Organization; 2004.
19. SNOMED International. SNOMED CT 2019. Available from: http://www.snomed.org/.
20. American Medical Association. CPT® (Current Procedural Terminology) 2019. Available from: https://www.ama-assn.org/amaone/cpt-current-procedural-terminology.
21. Regenstrief Institute. About LOINC 2019. Available from: https://loinc.org/about/.
22. Sagiroglu S, Sinanc D, editors. Big data: a review. In: 2013 International Conference on Collaboration Technologies and Systems (CTS). Piscataway: IEEE; 2013
23. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309(13):1351–2.
24. Chapman WW, Chu D, Dowling JN, editors. ConText: An algorithm for identifying contextual features from clinical text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Stroudsburg: Association for Computational Linguistics; 2007
25. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, et al. Toward better public health reporting using existing off the shelf approaches: a comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. J Biomed Inform. 2016;60:145–52.
26. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
27. Chen T, Guestrin C, editors. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016.
28. Évora L, Seixas J, Kritski AL. Neural network models for supporting drug and multidrug resistant tuberculosis screening diagnosis. Neurocomputing. 2017;265:116–26.
29. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology. 2017;2(4):230–43.
30. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature biomedical engineering. 2018;2(10):719–31.
31. Martin S, Wagner J, Lupulescu-Mann N, Ramsey K, Cohen AA, Graven P, et al. Comparison of EHR-based diagnosis documentation locations to a gold standard for risk stratification in patients with multiple chronic conditions. Appl Clin Inform. 2017;8(3):794.
32. Kasthurirathne SN, Biondich PG, Grannis SJ, Purkayastha S, Vest JR, Jones JF. Identification of patients in need of advanced care for depression using data extracted from a statewide health information exchange: a machine learning approach. J Med Internet Res. 2019;21(7):e13809.
33. Murray DD, Itenov TS, Sivapalan P, Eklöf JV, Holm FS, Schuetz P, et al. Biomarkers of acute lung injury the individualized approach: for phenotyping, risk stratification and treatment surveillance. J Clin Med. 2019;8(8):1163.
34. Fareed N, Swoboda CM, Jonnalagadda P, Griesenbrock T, Gureddygari HR, Aldrich A. Visualizing opportunity index data using a dashboard application: a tool to communicate infant mortality-based area deprivation index information. Appl Clin Inform. 2020;11(04):515–27.
35. Dixon BE, Grannis SJ, McAndrews C, Broyles AA, Mikels-Carrasco W, Wiensch A, et al. Leveraging data visualization and a statewide health information exchange to support COVID-19 surveillance and response: application of public health informatics. J Am Med Inform Assoc JAMIA. 2021;28(7):1363–73.
36. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020;20(5):533–4.
37. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574.
38. Hripcsak G, Schuemie MJ, Madigan D, Ryan PB, Suchard MA. Drawing reproducible conclusions from observational clinical data with OHDSI. Yearb Med Inform. 2021;30(1):283–9.
39. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The national COVID cohort collaborative (N3C): rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021;28(3):427–43.

40. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. In: Machine Learning for Healthcare Conference. Boston: PMLR; 2017.
41. Sorin V, Barash Y, Konen E, Klang E. Creating artificial images for radiology applications using generative adversarial networks (GANs)–a systematic review. Acad Radiol. 2020;27(8):1175–85.
42. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. eGEMs. 2016;4(1):1–21.
43. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit on Translational Bioinformatics. 2010;2010:1.
44. Ferryman K, Pitcan M. Fairness in precision medicine. Data & Society. 26 Feb 2018; Available from: https://datasociety.net/library/fairness-in-precision-medicine/. Accessed 8 Feb 2022.
45. Roebuck K. Data quality: high-impact strategies-what you need to know: definitions, adoptions, impact, benefits, maturity, vendors. London: Emereo Publishing; 2012.
46. Kim Y, Huang J, Emery S. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. J Med Internet Res. 2016;18(2):e41.
47. Shankar S, Halpern Y, Breck E, Atwood J, Wilson J, Sculley D. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:171108536. 2017.
48. Tommasi T, Patricia N, Caputo B, Tuytelaars T. A deeper look at dataset bias. Domain adaptation in computer vision applications. Cham: Springer; 2017. p. 37–55.
49. Buolamwini J, Gebru T, editors. Gender shades: intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency. Boston: PMLR; 2018.
50. World Health Organization. About social determinants of health 2020. Available from: https://www.who.int/social_determinants/sdh_definition/en/.