Big Data in Medical AI: How Larger Data Sets Lead to Robust, Automated Learning for Medicine

2

Ting Xiao and Mark V. Albert

2.1 Why the Big Data Revolution?

Machine learning is having a dramatic impact on the way we leverage information to make decisions [1, 2]. The success has been obvious in commercial business settings where data from advertising [3], supply logistics [4], and even social media [5, 6] is collected and processed in real time, enabling decisions at speeds and scales that would be impossible for hired employees. Medical applications present unique challenges due to risks but also provide satisfying targets due to the potential for improving health outcomes [7–10].

Many steps of the medical decision-making process can benefit from the tools of machine learning (Table 2.1). For example, we can consider a common sequence of choices made during the course of a medical treatment.

1. The clinician is tasked with collecting the relevant information.
2. A judgement about the cause is made based on the information available.
3. A treatment is proposed when possible.
4. Response to treatment is periodically evaluated and altered when needed.

T. Xiao
Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA

Department of Information Science, University of North Texas, Denton, TX, USA
e-mail: ting.xiao@unt.edu

M. V. Albert (✉)
Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA

Department of Biomedical Engineering, University of North Texas, Denton, TX, USA

Department of Physical Medicine and Rehabilitation, Northwestern University Feinberg School of Medicine, Evanston, IL, USA
e-mail: mark.albert@unt.edu

F. Jotterand, M. Ienca (eds.), *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues*, Advances in Neuroethics,
https://doi.org/10.1007/978-3-030-74188-4_2

Table 2.1 Definitions

Artificial intelligence (AI)	The development of computer systems performing tasks commonly associated with intelligent beings, either through explicit programming or by learning from data
Machine learning	A large subset of AI which makes data-driven inferences. Notably, this is the area in which the vast majority of AI advances are made
Big data	A term to describe the tools and techniques of inference that are particular to large data sets, which enable more robust, automated learning
Deep learning	Machine learning using multilayer ("deep") neural networks. Currently the state of the art in solving challenging inference problems with large data sets by learning intermediate features directly from raw data
TensorFlow, PyTorch	The two dominant deep learning frameworks
GPU	Graphics Processing Unit. A processor designed to handle graphics operations that can be used to dramatically speed up neural network training due to the similarly simple, distributed processing needs

Medical professionals are trained to perform each of these steps taking into account what they observe directly or measure, and they then relate that information to their own personal experience and the medical research. However, it is worth noting that each of these steps can loosely be associated with a related approach used in machine learning techniques which are particularly valuable for large data sets and suggest recommendations for complex decision-making problems. For example, here we can list four machine learning strategies that can be directly mapped to the four steps above to assist the clinician in certain cases:

1. *Feature selection*: With enough data, the process of determining which information is more or less important can be automated. If the data is difficult or invasive to collect, a ranking of the importance can be provided to help the clinician choose the best measures to collect for a diagnosis [11].
2. *Factor analysis*: Notwithstanding the philosophical arguments of truly establishing cause and effect relationships, much of approach to understand a collection of symptoms is finding the underlying factor or factors explaining the symptoms presented. This goes well beyond disease diagnosis. Underlying factors may be more fine-grained than disease states, or emerge from comorbid diseases—a factor analysis would be able to identify groups of common concern in an automated way to allow patients with similar conditions to be grouped and treated more effectively [12].
3. *Predictive modeling*: The choice of treatment relies on the belief of which option is expected to lead to the greatest improvement, while weighing appropriate risks. Clinical researchers use statistical models to evaluate the superiority of one treatment over another, and in ambiguous cases, medical practitioners also use internal estimates of future improvement through their years of medical experience. However, with larger data sets, such predictions can be explicit and even tailored to the particular hospital, patient group, clinician, surgical technique using available data on past outcomes to provide an additional point of reference to help make a treatment recommendation [13].

4. *Automated outcome data collection and synthesis*: For long-term treatments, follow-up is necessary to judge compliance, efficacy, and make adjustments as needed. However, visits to the clinic are costly in terms of clinician time and associated financial costs. Questions regarding symptoms in a clinical visit can be subjective or incomplete, and physical measures may differ based on a variety of factors. Sensor technologies exist now which enable convenient, continuous, and objective measures of a variety of symptoms, with associated analytics to distill the measures to clinically relevant information [14].

In short, machine learning, and the associated use of large data sets to improve the process of learning, can augment the process of clinical decision-making. Such analytics provide a unique perspective for each decision. Notably, such tools perform a similar function to a secondary consult or collective review among clinicians, without the associated time, costs, or overhead—enabling rapid, often automated assistance to inform medical care.

2.1.1 More Samples, More Features

One of the reasons for the explosion of machine learning is the availability of data for training decision-making systems. The amount of data varies along two dimensions that are particularly relevant to learning systems—additional samples and additional features. Samples generally represent more examples or cases. Features, on the other hand, are new types of information that can be collected for each sample. Modern technology has made it possible to dramatically increase both dimensions of data to build learning models. More data enable systems to be more capable of automated decision-making.

To understand why this is the case, let us begin with a common rule of thumb for collected data to train many standard machine learning prediction models.

$$n_{\text{samples}} \gg \left(n_{\text{features}}\right)^2$$

That is, the number of samples collected should be substantially greater than the square of the number of features. Double the number of features, and so the number of samples has to be quadruped, etc. Note this is only a rough "rule of thumb" with many exceptions. This is not as critical for some simpler prediction algorithms (such as Naive Bayes), but it is reasonably accurate for a number of common machine learning models which are sufficiently flexible and powerful to learn for a wider variety of prediction problems. Why is this true? That is beyond the scope of this chapter, but some motivation is provided in the footnote.[1]

[1] Succinctly, the goal of machine learning is roughly stated as the ability to group similar sample points together in a n_{features} dimensional space. Most ways of flexibly grouping points in a n-dimensional space require more than n^2 parameters (groups of planes, multidimensional ellipses, etc.), and a well-known fact of estimation is that you generally need more data points than you

The implication of a rule like this is that if there are not many samples, it is necessary to a priori select features for a learning model based on prior experience and intuition (which is often built based on prior experience); otherwise, there is not enough data for reliable learning. It is this limited hand-selection of features that generally leads to weaker performance when more data is available. If a massive amount of data can be collected prior to hand-selecting features, the process of selecting features becomes automated and in many cases more reliable than personal judgement.

In fact, the advantage of big data as a tool in medical decision-making primarily comes from this ability to automatically select, weigh, and combine features. Although many statistical tools have existed which can use features similarly, modern machine learning approaches built upon these techniques enable a slightly larger number of features relative to samples through a variety of strategies. Recall that including more relevant features universally improves predictive accuracy—including features that may seem irrelevant according to personal intuition. There are approaches which stray far from the approach of building a single model. We will discuss ensemble learning, which effectively polls disparate machine learning algorithms to arrive at an answer better than any algorithm alone [15]. Similarly, deep learning can combine the features present in raw data to create more complex features sets that can be leveraged to improve learning [16]. However, both ensemble learning and deep learning require a larger amount of data as they are hierarchical in nature needing well-trained component models to perform well.

Ultimately, however, even without these recent modeling advances, much of the improvement relies on the availability of larger data sets and computing systems capable of processing these vast quantities of data efficiently.

2.1.2 Hardware Improvements

The collection and processing of massive data sets has required new paradigms of data management. In industry, this has led to the creation of positions for data engineers whose primary role is to collect and manage the acquired data for later machine learning researchers and data scientists [17]. In addition to collection, the processing tasks are challenging with substantially more data requiring parallel architectures for processing, ranging from distribution between machine, cores, or even across GPUs for very low-level distributed processing. We will address each of these in turn.

Data collection has been made simpler and more standardized through enterprise data systems with shared resources. In commercial systems, this has occurred through large cloud-based data management architectures with scalable data repositories and shared processing repositories. To enable efficient use of centralized data management systems, virtualization has made it possible to access such systems

have parameters to estimate, suggesting most learning algorithms require significantly more than $(n_{\text{features}})^2$ samples.

with a variety of tools. Different operating systems, with different analysis tools and programming languages, can work in tandem on a shared repository of information, significantly speeding up adoption of centralized data management systems. Additionally, the ability to allocate a variable number of processors to tasks allows not only efficient data storage, but also processing resources which can scale larger for data mining and model learning tasks, and scale smaller for daily incremental development and deployment tasks.

High-performance computing has also enabled real-time processing tasks over terabytes of data. Architectures based on the concept of map-reduce (e.g., Hadoop, Spark) speed up queries through separating the computation on different processors and/or sections of the data set (map) and combining the results (reduce) [18, 19]. With a sufficiently resourced architecture, queries that traditionally would take hours or days are shortened to seconds. Tasks that involve preparing data for analysis, including data visualization and cleaning, are significantly sped up using such strategies.

One of the most recent advances in machine learning is the supremacy of deep learning for complex learning problems such as visual object recognition, speech recognition, and natural language processing. Deep learning neural network models are capable of learning directly from raw acquired data by building layers of features derived from earlier layers. This permits successively more complex feature extraction. However, due to the large number of learned parameters, deep learning neural networks are particularly resource intensive to train. Luckily, because the computations of individual neural elements are relatively simple, it is possible to distribute them among simpler processing units. As graphics processing traditionally relied on a large series of fast, simple linear computations, graphics processing units (GPUs) were built to efficiently distribute such low-level processing tasks. Deep learning neural network frameworks, such as TensorFlow and PyTorch, are able to shift processing from CPUs to GPUs with the speed of processing increasing by orders of magnitude.

These hardware advances enable large, centralized data repositories with shareable and configurable processor allocation. Standard data analysis tasks can be distributed among processors across terabytes of data in seconds speeding data preparation and standard analysis approaches. And finally, readily available GPU processing has enabled high-speed training of deep learning neural network models for progressively more challenging modeling tasks.

2.2 Precision Medicine

The traditional approach to medicine has been to treat patients in a similar manner to how all patients with similar causes or measured symptoms would be treated. However, this one-size-fits-all approach to medicine has been gradually replaced by precision medicine approaches which attempt to tailor prevention, diagnosis, treatment, and evaluation to the particular patient under consideration [20, 21].

On one extreme for preventative care, advances in genomics provided an idealized case of data-driven medical intervention. For example, by identifying particular genetic anomalies, cancer risks can be assessed, and patients can make informed decisions to minimize the chance of developing a cancer before any symptoms have been identified. However, lifestyle analytics provide an alternate extreme where the data may be noisy and causal inferences unclear, but the resolution of advice about changing lifestyle may permit improved health outcomes on a massive scale—for example, precisely determining what quantity and manner of exercise is ideal for each person given their health needs, career requirements, and compliance concerns.

Precision diagnosis and treatment can rely on more than simply demographic information to increase or decrease the probability of a particular diagnosis. Clinicians are aware that individuals respond to different treatments depending on aspects of their physiology, mental health, compliance, and lifestyle. Instead of relying on clinician judgement, given these factors to influence how treatment choices are selected and presented to the patient, predictive models can be provided additional information on the state of the patient to systematically rank potential treatments.

Precision evaluation and follow-up depend upon reliable, readily acquired information on a subject's well-being. Reliability can be obtained by more frequent, more objective measures. Such information can often be acquired by passively worn wearable devices that can measure movements [22]; these devices can range from simple consumer step counters and calorie estimators [23–25] to research-grade wearables [26, 27] or even smartphones [28–31]. The interpretation of such movements to clinically relevant activities can often require analytics tailored to the individual population movements [32, 33] and potentially even the context such as at-home versus in the clinic [27, 34]. The convenience of wearable devices and other passively recorded health outcome measures has a definite impact on the future of medicine.

2.2.1 Challenges and Opportunities Unique to Mental Health and Wellness

Mental health has particular challenges in precision medicine. The first distinction from physical health is the lack of straightforward physical metrics to measure. For example, someone who suffers from depression can readily list a variety of symptoms which can be used to make a clinical diagnosis; however, affixing a sensor to them to collect data to make as certain of a diagnosis would be challenging.

One way of addressing limited mental health measures is to provide ways of allowing people undergoing treatment to periodically self-assess. However, this suffers from a number of drawbacks. First, questions require self-reflection or a recall of past experiences, both of which are known to be error, subjective and prone to error, or misrepresentation [35]. Additionally, compliance can be challenging, particularly in the case of individuals whose desire to respond regularly is affected by the very mental state that is being assessed.

Wearable device analytics paired with predictive methods on summary metrics prove a potential avenue for real-time assessment. For example, subjects suffering from depressive episodes may move less, speak less, and engage in more passive forms of entertainment. Any individual measure alone may not provide sufficient information for an estimate that a depressive episode occurred, however, taken in aggregate in the form of a prediction algorithm they would provide a reliable way of probabilistically estimating a depressive episode occurred. This could provide real-time scoring and large-scale data collection to assess the efficacy of group therapies and construct interventions particular to individual populations.

2.3 Tools for Big Data in Medicine

2.3.1 Standardization Tools

One of the greatest impairments in large-scale machine learning projects is the capture of data in formats readily available for analysis. This is particularly problematic in medical contexts where clinicians in different hospitals may record information in ways which are not compatible—for example, using different metrics or different units to quantify a given symptom. Even with the same type of data, the electronic systems may use different data formats, query languages, or permissions. This heterogeneity limits the application of big data for solving more challenging problems in medicine.

There are different approaches to standardizing data sets. One is to use a shared vocabulary for acquired data. For example, the International Classification of Diseases (ICD) and the Systematized Nomenclature of Medicine and Clinical Terms (SNOMED CT) are available standards [36, 37]; however, these are limited to fairly narrow data sets and terminology. More expansive common data models are being developed, including the Observational Medical Outcomes Partnership (OMOP) [38]. Notably, the standardization of data fields is not only helpful for big data analytics approaches. Standardization also enhances the ability for models to be trained, validated, and tested in different settings, increasing the level of scientific rigor possible prior to standardization. In addition to shared standards, shared access to computational tools for data storage, queries, and analytics support the use of centralized access of the data collected across institutions. This standardization of data and analytics tools enables rapid development and testing of predictive models on acquired health data.

2.3.2 Analytics to Leverage Big Data

A variety of machine learning techniques are available to exploit collected data for improved understanding; however, a subset of them have been particularly valuable in making inferences on large, often noisy data sets as is typical in medical contexts.

2.3.2.1 Unsupervised and Semi-Supervised Learning Methods

Measuring the extent of diseases or disease progression, particularly with mental health, can be challenging. Many more types of information can be collected relative to the number of individuals participating in the study—this is particularly the case with continuously collected observational data (e.g., from wearable devices). As discussed previously, in order to make valid inferences, the number of features must be reduced when fewer samples are available.

One approach to shrink the number of features in a learning model is to synthesize a large number of features into fewer, more reliable aggregate factors. Principal components analysis (PCA) and related factor analysis techniques provide one means of achieving this; however, a variety of methods now exist to parameterize a larger data set with a large number of features into a smaller, more meaningful set of features. Some are based on analytical assumptions about the distribution of underlying causes, such as with PCA or ICA (independent components analysis), while other reductions are made possible through a nonparametric combination of features using the hidden layers of neural network models to effectively compress the data.

Additionally, one of the challenges in building machine learning models in medicine is properly vetting the quality of decisions the model is using for learning. For example, some scoring methods may have low inter-rater reliability and require group consensus for high reliability. When high-quality labeled data is scarce for training, but unlabeled collected data is plentiful, there are two particular machine learning strategies that work well to exploit the large amount of unlabeled data—anomaly detection and semi-supervised learning. If the positive diagnoses are exceptionally rare, and are the main source of "unusual" observations, anomaly detection techniques use the deviation from typical to better identify potential future cases. However, if there are multiple classes and there is sufficient labeled data for all classes being considered, semi-supervised learning techniques provide a means of estimating labels for all samples. In label spreading, a standard semi-supervised learning technique, the unlabeled samples that are easiest to classify are labeled first until all samples have been classified. In general, this allows clusters around labeled values in the feature space to be grouped according to similarity, giving standard machine learning models access to larger, labeled data sets.

2.3.2.2 High-Throughput Model Selection and Testing

A standard approach to statistical modeling involves selecting the features and statistical model prior to analysis—this greatly simplified the problems inherent in iteratively trying many different models when making statistical inferences. However, now there are a wide variety of models with many modeling parameters and options, which can all be trivially tried and adapted to each data set by altering, in many cases, only a single line of code. For example, the following variations can be easily attempted to fit the best model:

1. Changing the input features through feature selection or feature engineering (e.g., dimensionality reduction or clustering).

2. Setting hyperparameters of a model (e.g., the degree of a polynomial, the regularization strength of lasso regression).
3. Selecting different models (e.g., support vector machines vs. random forest).
4. Combining models through ensemble learning (more in the next subsection).

If we systematically test all these variants on a single set of data, the best fitting model would likely overfit. Generally, the most complex models can readily fit a given data set for training, but without proper tuning to prevent overfitting, they would perform poorly on newly acquired samples.

A regiment of separating the data sets into training, validation, and testing sets is standard practice in machine learning, and critical for proper selection among all these alternate models and parameterizations. This will be addressed further in Sect. 2.4.1.1.

Fortunately, each combination of modeling options can be tested independently. The space of options can be explored rigorously through a grid search or can be optimized adaptively using a variety of available optimization strategies based on model performance. Different model combinations can readily be tasked to different CPU cores, greatly increasing the opportunity to tune a variety of models for a given predictive problem.

2.3.2.3 Ensemble Methods

In predictive analytics competitions, it is not uncommon for competing groups to "team up" to improve their predictive accuracy beyond either classifier alone using ensemble methods. This is a common enough strategy that many competitions ban the practice; however, if the goal is to increase prediction quality, this is precisely a strategy worth exploring. Models which may bear no resemblance to each other analytically can combine their results to form a prediction which is generally at least as good as the best individual model. In fact, this ensemble approach is the basis for a very effective predictive model, random forest, which is aptly built from the systematic combination of predictions from individual decisions trees.

2.3.2.4 Deep Learning

Generally, machine learning algorithms require designers to extract features from the raw signals using their intuition or previous experience to identify features where a fair amount of the extracted features are at least weakly capable of assisting classification. For many problems, this procedure works well—it is the standard means of generating predictive models for hundreds of years. However, this requirement for human implementation and selection of features is a major impediment in many machine learning tasks. Many problems require a complex series of intermediate feature combinations which are not readily built by human designers. For example, object identification occurs in the brain through a series of processing stages moving from simple features to complex feature combinations. Similarly, the best performing visual object identification systems also process visual information through a series of layers of processing of artificial neurons. The intermediate features are often impossible to describe plainly, but such data-driven low-level

features are essential for the algorithm to make distinctions in rich, complex, raw data sets.

Currently, the field of machine learning has developed many tools related to deep learning analytics to speed programming and processing. TensorFlow and PyTorch are the de facto standards of frameworks for coding and training deep learning neural network models. By creating models in TensorFlow or PyTorch, designers also benefit from an ability to readily power their models to run on available GPUs for significant speed improvements.

Ultimately, these data management and analytic tools provide a great deal of power to model designers, particularly for those with access to large data sets.

2.4 Big Data Concerns

Medicine presents a unique set of challenges in the development of predictive models. There are high stakes in assuring that model predictions not only are accurate in the original context but also function robustly in other contexts. In some cases, lives literally depend on high accuracy or robust behavior. Beyond concerns of predictive ability, the medical data necessary to make such predictions is particularly sensitive and personal, requiring extra care to keep it secure. And finally, even with accurate models and adequate precautions for individual data security, the results of predictive models can be as biased as the human medical decisions on which they are based—care must be taken to avoid systematic bias and identify when particular populations may not be adequately represented which affects accuracy as well as the potential for bias.

2.4.1 The Need for Proper Validation

Accuracy is paramount in medical decision-making; however, there are a variety of factors in machine learning which can affect the accuracy of real-world applicability of machine learning models. We address each major concern in turn.

2.4.1.1 Test Models with Separating Training, Validation, and Test Sets

When a large number of models are tested, it becomes more critical to have separate training and test sets to avoid overfitting which causes both inflated expected performance and the selection of models which perform poorer on new data sets. However, when an exceptionally large number of models are iteratively tested using a validation set (a test set used for model selection), as is the case with modern automated machine learning model fitting and selection procedures, it is critical to properly evaluate any selected model on an independent test set. This is easily automated with sufficient data availability, though many cases of applied machine learning models do not evaluate their models with proper test sets after model selection. We recommend anyone using machine learning approaches to not only consider

traditional cross-validation in model selection with a hold-out data set, but also consider nested cross-validation techniques to avoid overfitting and reporting accurate quality metrics for the selected, trained models.

2.4.1.2 Assure Samples Adequately Represent the Reported Populations and Contexts

Many researchers are aware of the need to have an appropriate representation of sexes and races for their results to generalize across the population; however, there are a number of other factors to consider that can be equally important for proper statistical representation. For example, it may be necessary to design not only a model for a particular disease population but also a subpopulation of that group with particular characteristics or comorbidities. Similarly, if assessing behavior, it may be necessary to acquire data in particular contexts, for example, not only in the clinic but also at home.

2.4.1.3 Avoid Contamination Between Training and Test Sets

For proper model validation, the test set should be independent of the training and validation sets to avoid inflating reported model accuracy. However, this understood rule can often be broken without realizing it. One common example of this is having the same subject's data in both the training set and test set. Even if the data samples are independent, the individual may introduce a dependency between samples that promote overfitting. This may not be a concern with a model that will be trained with an individual's own data for personal use, but most models in medical practice are trained, fixed, and deployed without adaptation to a single individual. For this reason, training and test sets should involve separate individuals—commonly performed using subject-wise cross-validation. This often leads to lower predictive accuracy which makes it less palatable of a result to emphasize in publication, but provides a much more realistic assessment of models in a deployed context.

2.4.1.4 Select the Right Performance Metrics

The need for proper metrics to evaluate models is particularly strong in medical context and, without proper consideration metrics, can be selected by marketers of prediction systems that can border on fraud. For a classic example, a system can report a 99% accuracy for tumor detection in radiological scans, but fail to indicate the model just naively reports for every scan that no tumor is present, which is accurate for 99% of the population.

Various metrics exist to tease apart the behavior of a system for a clearer estimate of efficacy. The classic metrics for binary classification are sensitivity and positive predictive value. However, given the increase of multiclass classification problems, an introduction and emphasis of recall and precision, their multiclass equivalents, may be more appropriate. In cases where it is unclear whether recall or precision is favored, F1 score is the harmonic mean of precision and recall. By averaging across all classes, average F1 score provides a metric more in line with a sense of practical efficacy.

In regression, many models are designed to optimize mean squared error by default. However, there are a variety of other metrics which may be more appropriate. Mean absolute error is more forgiving of errors in outliers to better minimize more moderate errors. Additionally, there are log error metrics which are best to minimize the relative or percent error rather than the absolute scale of errors. For example, a model predicting medical care costs across a wide range of scales would likely benefit from using log error rather than mean squared error. A $1000 error in an expensive surgery should be penalized less than a $1000 error in low-cost routine care, whereas a 1% error for the surgery may be comparable to a 1% error in routine care depending on the use case of the cost of care model. The key point is that it may be important to select an appropriate metric rather than rely on defaults, as all metrics make assumptions either explicitly or implicitly about which errors are more important to improve the model.

2.4.2 Security

Models in medical decision-making often rely on sensitive information that is protected by various regulations including HIPAA, the Health Insurance Portability and Accountability Act in the United States [39], and GDPR, the General Data Protection Regulations in the European Union [40], which restrict the ways in which health information can be shared. Primary among the protocols for data sharing are the information is inaccessible outside the healthcare provider or intended original use, and the assumption that analytics and results derived from this data do not adversely affect an individual.

For data collected in a hospital setting and saved on centralized hospital servers, standard data security concerns are applicable with protocols vigilantly adhered to in order to avoid data security breaches. However, as we enter an era of wearable devices, phone-acquired data, smart home health, etc., the constellation of "Internet of Things" data security can be more challenging. This is particularly relevant as users may not be aware of all the inferences that can be made from data they consider innocuous. For example, on a mobile phone alone, the accelerometer can be used to not only estimate step counts and calories but also predict personal activities such as trips to the bathroom or bedroom sleeping habits. For this reason, it is important that all information used for medical decision making be collected securely and encrypted not only when stored locally but also en route to a centralized repository.

2.4.3 Ethical Challenges

Hardware and analytics advances have led to the potential for powerful and accurate automated decision-making; however, there are problems that more powerful prediction algorithms may not only avoid addressing but also make far worse. It is important to take the broader context of predictive algorithms into account. For example, imagine one group of patients is diagnosed improperly in a systematic way (e.g., based on race, sex), which can often occur when decision-making is done

by humans with their own implicit biases. Due to the nature of machine learning algorithms, the training data in this case will also lead to systematic errors in this group of patients. One might naively assume that if we simply remove the label of that group from the features a machine learning system can use, that might solve the potential for discrimination, but that is not true. Even without use of the group label during training or application of a model, or even if it is not collected during the training phase, a developed system can become biased against that group by propagating the poor choices that were made in the original data set.

Although humans can report on aspects of their decision-making, and we have developed relatively sophisticated means of establishing if someone is biased through a combination of nonverbal cues, past decisions, and history, the same level of insight is more challenging for automatic decision systems. This is particularly true for deep learning neural networks which, though quite powerful for complex decision-making, are also more challenging to explain the steps of decision-making. Tools are available to observe bias in automated decision-making systems, but this requires additional data collection and discussion of fairness using statistical arguments that may not be standard practice in machine learning or clinical teams.

2.5 Conclusion

Medical decision-making has relied on collected evidence and experience to generate models to predict outcomes and influence the choice of treatment options. Until relatively recently in history, this process has been limited to human experience and decision-making; however, the advent of technologies to collect and curate massive data sets has led to new capabilities not previously possible. Consumers are observing cars that can drive themselves by identifying objects in real-time at high speeds, systems that can generate narrative prose and speech which sounds natural, and systems that regularly win at strategic games like chess or go. Medicine is currently embracing the same big data technologies that enable these feats that were previously thought unique to human capability and experience. By leveraging the larger data sets available, systems can learn to pay attention to relevant features, weigh them, and combine them systematically to arrive at more accurate decisions than human decision-making alone. This will not only aid clinical research to help provide human-interpretable insights to improve standard clinical care as it is practiced today, but it will increasingly be integrated in an automated way to clinical care and early prevention and screening, dramatically shaping the practice of medicine going forward.

References

1. Angra S, Ahuja S. Machine learning and its applications: a review. In: 2017 international conference on big data analytics and computational intelligence (ICBDAC). 2017. p. 57–60.
2. Portugal I, Alencar P, Cowan D. The use of machine learning algorithms in recommender systems: a systematic review. Expert Syst Appl. 2018;97:205–27.

3. Juan Y, Lefortier D, Chapelle O. Field-aware factorization machines in a real-world online advertising system. In: Proceedings of the 26th international conference on world wide web companion. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE. 2017. p. 680–8.
4. Wenzel H, Smit D, Sardesai S. A literature review on machine learning in supply chain management. In: Artificial intelligence and digital transformation in supply chain management: innovative approaches for supply chains. Proceedings of the Hamburg international conference of logistics (HICL), vol. 27. Berlin: epubli GmbH; 2019. p. 413–41
5. Romanov A, Semenov A, Mazhelis O, Veijalainen J. Detection of fake profiles in social media - literature review. In: Proceedings of the 13th international conference on web information systems and technologies. https://doi.org/10.5220/0006362103630369.
6. Singh J, Singh G, Singh R. Optimization of sentiment analysis using machine learning classifiers. HCIS. 2017;7:32.
7. Zhou H, Tang J, Zheng H. Machine learning for medical applications. ScientificWorldJournal. 2015;2015:825267.
8. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics. 2017;37:505–15.
9. Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, Brink J. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. J Am Coll Radiol. 2018;15:504–8.
10. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
11. Gronsbell J, Minnier J, Yu S, Liao K, Cai T. Automated feature selection of predictors in electronic medical records data. Biometrics. 2019;75:268–77.
12. Caballero FF, Soulis G, Engchuan W, Sánchez-Niubó A, Arndt H, Ayuso-Mateos JL, Haro JM, Chatterji S, Panagiotakos DB. Advanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: the ATHLOS project. Sci Rep. 2017;7:43955.
13. Tang F, Xiao C, Wang F, Zhou J. Predictive modeling in urgent care: a comparative study of machine learning approaches. JAMIA Open. 2018;1:87–98.
14. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Per Med. 2018;15:429–48.
15. Zhang C, Ma Y. Ensemble machine learning: methods and applications. Boston, MA: Springer; 2012.
16. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
17. Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making. Big Data. 2013;1:51–9.
18. Zaharia M, Xin RS, Wendell P, et al. Apache spark: a unified engine for big data processing. Commun ACM. 2016;59:56–65.
19. Shvachko K, Kuang H, Radia S, Chansler R. The hadoop distributed file system. In: 2010 IEEE 26th symposium on mass storage systems and technologies (MSST). ieeexplore.ieee.org. 2010. p. 1–10.
20. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. BMC Med Inform Decis Mak. 2018;18:139.
21. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372:793–5.
22. Albert MV, Shparii I, Zhao X. The applicability of inertial motion sensors for locomotion and posture. Locomotion and posture in older adults. 2017.
23. Qiu S, Cai X, Chen X, Yang B, Sun Z. Step counter use in type 2 diabetes: a meta-analysis of randomized controlled trials. BMC Med. 2014;12:36.
24. Modave F, Guo Y, Bian J, Gurka MJ, Parish A, Smith MD, Lee AM, Buford TW. Mobile device accuracy for step counting across age groups. JMIR Mhealth Uhealth. 2017;5:e88.
25. Albert MV, Deeny S, McCarthy C, Valentin J, Jayaraman A. Monitoring daily function in persons with transfemoral amputations using a commercial activity monitor: a feasibility study. PM R. 2014;6:1120–7.

26. Albert MV, Sugianto A, Nickele K, Zavos P, Sindu P, Ali M, Kwon S. Hidden Markov model-based activity recognition for toddlers. Physiol Meas. 2020;41:025003.
27. Albert MV, Azeze Y, Courtois M, Jayaraman A. In-lab versus at-home activity recognition in ambulatory subjects with incomplete spinal cord injury. J Neuroeng Rehabil. 2017;14:10.
28. Shawen N, Lonini L, Mummidisetty CK, Shparii I, Albert MV, Kording K, Jayaraman A. Fall detection in individuals with lower limb amputations using mobile phones: machine learning enhances robustness for real-world applications. JMIR Mhealth Uhealth. 2017;5:e151.
29. Antos SA, Albert MV, Kording KP. Hand, belt, pocket or bag: practical activity tracking with mobile phones. J Neurosci Methods. 2014;231:22–30.
30. Albert MV, Kording K, Herrmann M, Jayaraman A. Fall classification by machine learning using mobile phones. PLoS One. 2012;7:e36556.
31. Albert MV, McCarthy C, Valentin J, Herrmann M, Kording K, Jayaraman A. Monitoring functional capability of individuals with lower limb amputations using mobile phones. PLoS One. 2013;8:e65340.
32. Albert MV, Toledo S, Shapiro M, Kording K. Using mobile phones for activity recognition in parkinson's patients. Front Neurol. 2012; https://doi.org/10.3389/fneur.2012.00158.
33. Sok P, Xiao T, Azeze Y, Jayaraman A. Activity recognition for incomplete spinal cord injury subjects using hidden markov models. IEEE Sensors J. 2018;
34. O'Brien MK, Shawen N, Mummidisetty CK, Kaur S, Bo X, Poellabauer C, Kording K, Jayaraman A. Activity recognition for persons with stroke using mobile phone technology: toward improved performance in a home setting. J Med Internet Res. 2017;19:e184.
35. Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol. 1990;43:87–91.
36. Bowman SE. Coordination of SNOMED-CT and ICD-10: getting the most out of electronic health record systems. Coordination of SNOMED-CT and ICD-10: getting the most out of electronic health record systems/AHIMA, American Health Information Management Association. 2005.
37. Nosek BA, Alter G, Banks GC, et al. Scientific standards. Promoting an open research culture. Science. 2015;348:1422–5.
38. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, Welebob E, Scarnecchia T, Woodcock J. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010;153:600–6.
39. Annas GJ. HIPAA regulations - a new era of medical-record privacy? N Engl J Med. 2003;348:1486–90.
40. Tovino SA. The HIPAA privacy rule and the EU GDPR: illustrative comparisons. Seton Hall Law Rev. 2017;47:973–93.