Neurotherapeutics (2023) 20:1066–1080
https://doi.org/10.1007/s13311-023-01384-2

CURRENT PERSPECTIVES

Machine Learning in Clinical Trials: A Primer with Applications to Neurology

Matthew I. Miller[1] · **Ludy C. Shih**[2] · **Vijaya B. Kolachalama**[1,3]

Accepted: 21 April 2023 / Published online: 30 May 2023

Abstract

We reviewed foundational concepts in artificial intelligence (AI) and machine learning (ML) and discussed ways in which these methodologies may be employed to enhance progress in clinical trials and research, with particular attention to applications in the design, conduct, and interpretation of clinical trials for neurologic diseases. We discussed ways in which ML may help to accelerate the pace of subject recruitment, provide realistic simulation of medical interventions, and enhance remote trial administration via novel digital biomarkers and therapeutics. Lastly, we provide a brief overview of the technical, administrative, and regulatory challenges that must be addressed as ML achieves greater integration into clinical trial workflows.

Keywords Machine learning · Clinical trials · Neurology

Introduction

The term artificial intelligence (AI) refers to the use of computational methods to enable machines to perform tasks such as perception, reasoning, learning, and decision-making. Advances in the technology sector are fueling the development of novel forms of AI, which are rapidly driving progress across diverse domains such as facial recognition, financial strategy, and self-driving vehicles [1, 2]. The field of medicine is no exception, with AI methods increasingly being applied in healthcare research, from the laboratory to the bedside. In clinical trials, particularly, automated methods similarly carry great promise to alleviate many of the considerable difficulties associated with planning, completing, and analyzing the results of large scale trials. The challenges associated with traditional trials, from recruiting participants across diverse populations to the selection of feasible and appropriate eligibility criteria, make these interventions an ideal area for the application of emerging data science techniques.

In this article, we reviewed machine learning (ML) as a means of achieving AI and improving the practice of clinical research. We provided a basic introduction to key ML concepts for clinicians, surveyed general areas of application for ML in clinical trials, and then demonstrated how ML is being used to foster innovation in clinical research for neurologic diseases, specifically. We concluded with a discussion of technical challenges to automation in trials, highlighting potential obstacles that must be overcome to sustain innovation in the field.

Background

ML in Medicine: Why Now?

Efforts to standardize clinical care via advanced statistical models have their roots in the twentieth century [3, 4], when the advent of modern computers enabled researchers to begin simulating the process of differential diagnosis [3-8], recommending antibiotic regimens [9], and identifying medication effects [10]. Though these early initiatives fell short of making widespread impact [11], a number of factors have led to an unprecedented rate of progress in ML since the early 2010s.

Increased access to large quantities of electronic data (in medicine, most notably, publicly available datasets such as the UK Biobank [12] and the Cancer Genome Atlas [13]),

✉ Vijaya B. Kolachalama
vkola@bu.edu

[1] Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, 72 E. Concord Street, Evans 636, Boston, MA 02118, USA

[2] Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA 02118, USA

[3] Department of Computer Science and Faculty of Computing & Data Sciences, Boston University, Boston, MA 02115, USA

advances in computer hardware (especially Graphics Processing Units [GPUs]), and the widespread availability of open source software [14] have created the necessary environment for AI to achieve significant gains. Furthermore, continued algorithmic developments have enabled machines to take on tasks of increasing complexity and nuance [15].

Recent advances in machine learning have been driven by the development of novel techniques that prevent overfitting [16-20], and improve training processes [21-23], leading to the maturation of the field. Modern deep learning frameworks such as convolutional neural networks (CNNs) have emerged as a powerful tool for computer vision tasks [24], enabling the extraction of salient visual features from natural and medical images without the need for manual intervention. In addition, the development of new "transformer" networks has revolutionized machine learning models' ability to make context-aware predictions [25]. Overall, these advances have significantly improved the performance and versatility of deep learning in a range of applications. As a result, we have seen dramatic improvements in areas as diverse as speech recognition, driverless cars, and precision marketing of advertisements [26]. Medical innovation often follows directly from the progress made by software companies in non-clinical arenas [27], and healthcare researchers are increasingly using ML methods to augment clinician workflow, predict outcomes, and discover insights from medical datasets. From the accurate diagnosis and classification of skin cancer [28] to AI-based detection of diabetic retinopathy [29] to the potential for timely identification of Alzheimer's disease using both neuroimaging and clinical data [30], medical ML is showing its prowess to provide high-value contributions to patients and clinicians.

How Machines Learn: What Clinicians Should Know

While the notion of learning implies some measure of human-like agency, medical ML algorithms depend on the transformation of patient-derived data into numerical formats that can be processed by computer systems. For instance, computed tomography (CT) scans can be understood as matrices of pixel intensities, and vital sign measurements may be translated into lists or vectors of discrete measurements. If an investigator can derive numerical quantities from a given data source, then the possibilities for which modalities can be used as input to an ML strategy are nearly limitless.

With the data thus translated, ML models act according to principles encoded within their architecture. Supervised learning models, as an example, are traditionally composed of models that can be trained by minimizing an error, via a loss function, between their predictions and known quantities within a dataset that are typically provided by a human labeler [1, 2]. The loss function guides the model by adjusting its underlying mathematical structure (i.e., the parameters that govern the mappings from inputs to outputs) [31] so that the model can ultimately provide as output either a probabilistic estimate of a data point belonging to a certain category (in the case of classification tasks) or direct estimates of a continuous measurement (in the case of regression tasks). Nevertheless, the traditional paradigm of minimizing loss with human-supplied labels for prediction is increasingly in flux. Self-supervised learning models are coaxed to identify common patterns in data by being trained to associate samples with certain characteristics, such as those from the same source [32] (e.g., serial ECGs from a single patient). These models undergo "pretext" training to learn these associations without requiring explicit supervision and can then be repurposed for other tasks down the line, such as prediction. Reinforcement learning (RL) models, on the other hand, respond to "rewards," which direct the model into adjusting its parameters such that it increases its probability of performing certain actions [33] (e.g., making appropriate decisions in response to sepsis in intensive care settings) [34]. Additionally, generative models produce novel data products from either structured inputs that are then enhanced in some way (e.g., production of high-resolution radiologic scans from low-resolution analogs) [35] or even from simply statistical noise [36].

While different ML algorithms carry their own sets of advantages and disadvantages, the choice of which to use may depend on the task of interest, the available data, access to proper computing hardware, and the investigator's desire to elucidate mechanistic insights (i.e., interpretability) from the model. As an example, CNNs perform excellently in determining diagnoses from radiologic images. However, such models often contain millions of parameters, and when run on standard "central processing units" (CPUs), they are prohibitively slow to train and develop in iterative fashion. Specialized hardware, such as GPUs, are often needed to accelerate the pace of computation to a tractable timeline [37], but may not be as easily accessible in many environments. Logistic regression, on the other hand, may require little more than a desktop computer while yielding mathematical coefficients that can be intuitively interpreted in the context of the underlying data. Furthermore, complex models and ever-increasing amounts of data do not necessarily translate to higher performance. Simple data distributions (e.g., finding a best-fit line in a unidimensional scatterplot) do not require complex model architectures for adequate solutions to be discovered; indeed, in certain instances, simpler models may be found to perform near-equivalently to complex ones after comparison [38].

Lastly, the performance of medical ML models can be assessed according to a variety of metrics, depending on the specific use cases. In the case of diagnostic or prognostic classification tasks, it is often standard to report area under the receiver operating characteristic curve (AUROC), obtained by plotting true positive rate versus false positive rate at differing probability thresholds when comparing predictions versus observation [39]. Area under the precision-recall curve (AUPR) (obtained from plotting positive predictive value versus sensitivity) may also be reported, as AUROC may overestimate performance in the case of highly imbalanced datasets [40]. A variety of specialized metrics for tasks such as segmentation (e.g., dice coefficient and intersection-over-union) [41], image generation (e.g., structural similarity) [42], and other tasks may also be deployed depending on the use case. Conversely, in regression for continuous quantities, standard metrics such as the mean squared error (MSE) between predicted and observed values may also be used [43]. Regardless of the specific measure employed, however, it is also imperative that ML models be judged in terms of traditional criteria (e.g., sensitivity, specificity, accuracy) in order to fully contextualize their impact on patient care prior to deployment. An overview of essential ML terminology along with definitions is provided in Table 1. Examples of widely used ML algorithms are illustrated in Fig. 1 and further elaborated in Table 2.

Learning point 1: Machine learning frameworks have the potential to accelerate the timeline of clinical trials by facilitating patient selection via mining electronic health records.

AI and Clinical Trials

What Can be Gained?

Despite their successes, modern clinical trials remain difficult for research teams to bring to completion. Remarkably, unsuccessful trials remain the norm rather than exception due to myriad difficulties in identifying, enrolling, and providing treatment to patients within RCTs. Indeed, it has been estimated that only 12% of drug development programs achieve clinical trial success from phase 1 to launch [59]. While lack of clinical efficacy makes up a large component of the failures, many clinical studies fall short of recruitment goals and timelines due to factors such as low patient participation in clinical research and overly stringent inclusion criteria [60].

In what ways, then, can ML technologies help to alleviate these difficulties and advance new generations of clinical research? Here, we review several key areas in which such progress is already being demonstrated. We begin by discussing the power of natural language processing (NLP) approaches for sorting through large administrative databases and easing the work of identifying and screening potential participants. We next turn our attention to emerging methods for ML-based simulation of treatment interventions, which may one day challenge the supremacy of centralized, prospective studies. Lastly, we examine the possibility of medical software whose goal is not to support existing treatments but rather to act as the treatment in and of itself. These "digital therapeutics" require a rethinking of both the nature of medical therapy as well as the regulatory

Table 1 Basic machine learning nomenclature

Machine learning (ML): The study of statistical models with the capacity to improve their predictive performance with exposure to data
Classification: The task of using ML models to predict categorical labels from data e.g., predicting a diagnosis of Alzheimer's disease from an MRI scan
Regression: The task of using ML models to predict continuous labels from data e.g., predicting a neurocognitive test score from an MRI scan
Loss function: A metric that quantifies an ML model's prediction error. Over the course of training, the model "learns" to improve its predictions by minimizing the loss function
Training set: The dataset of examples that an ML model uses to minimize its loss function. Therefore, this is the set of all observations with which the model is "trained" to detect patterns in data e.g., a retrospective dataset of case and control patients was used to train a model that diagnoses Parkinson's disease
Testing set: The dataset to which a fully trained ML model is applied. This dataset is used to gauge the ability of the ML model to function in making real-world predictions e.g., a prospective patient population in which a newly trained model will be used to diagnose Parkinson's disease
Supervised learning: A subtype of ML in which models learn to make predictions by minimizing the error between their predictions and a set of predetermined outcomes (otherwise known as labels) e.g., teaching a model to diagnose ocular palsies by exposing it to a training set consisting of pre-labeled videos of cranial nerve examinations
Unsupervised learning: A subtype of ML in which models learn to infer patterns without being guided by predetermined labels. It does not require extensive manual labeling of data by human workers e.g., hypothesizing new subtypes of nonconvulsive status epilepticus from unlabeled electroencephalogram recordings
Semi-supervised learning: A hybrid approach spanning supervised and unsupervised learning. With this approach, an ML model learns to make predictions on large quantities of unlabeled data using a smaller labeled dataset as a "guide" e.g., teaching a computer vision model to highlight areas of all areas of hemorrhage in a head CT when only several axial slices have been annotated by a neuroradiologist

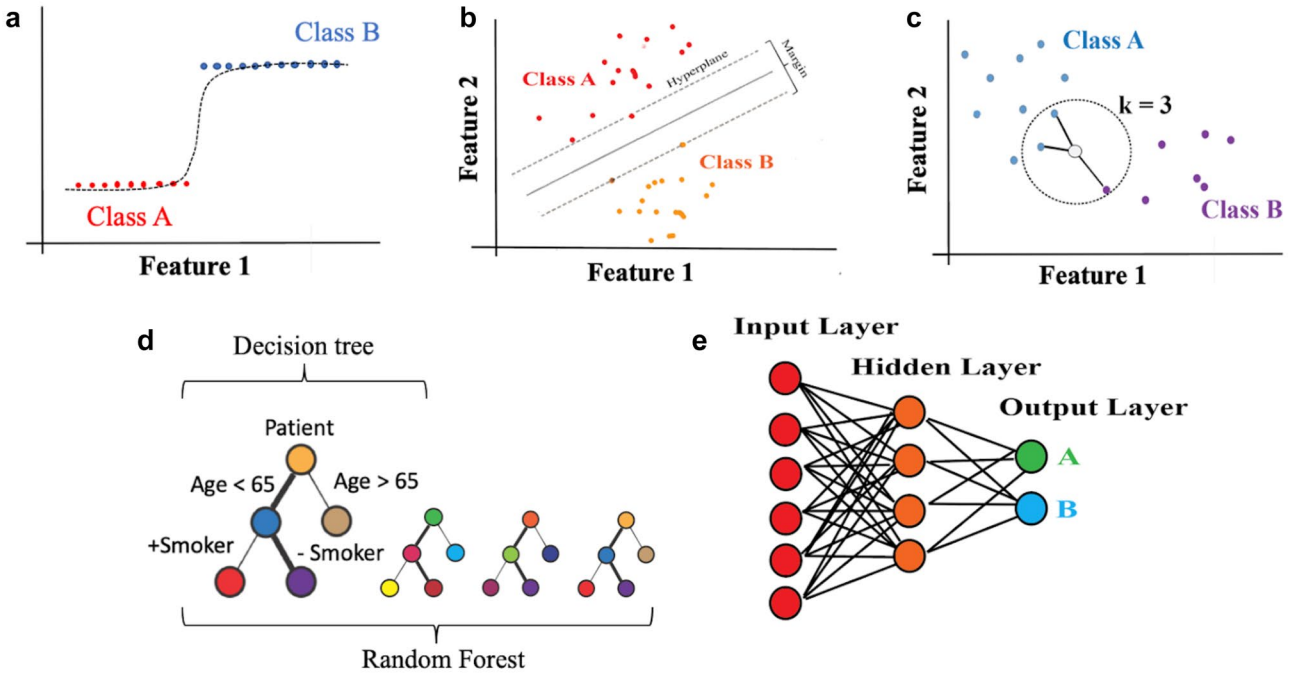

Fig. 1 Graphical illustration of machine learning algorithms. Schemata of several exemplary ML algorithms are demonstrated. **a** Logistic regression; **b** support vector machine (SVM); **c** K-nearest neighbors (KNN); **d** decision trees and random forests (RF); **e** neural network (specifically, a multilayer perceptron/MLP)

processes that govern the development and approval of drugs and devices.

Clinical Trial Execution: Patient Recruitment and Eligibility Criteria

A uniform problem across industry-, foundation-, and federally- funded clinical trials is their significant financial costs and lengthy timelines. Recent surveys of phase 3 studies, for instance, have demonstrated median durations of more than 700 days between the initial planning of a study and its initiation [59], while the expense of recruiting patients meeting eligibility criteria consumes ~ 1 billion US dollars in annual research spending and up to 30% of development timelines [61]. Indeed, identification of study participants rather than the conduct of the trial itself currently accounts for some of greatest timeline delays. Furthermore, despite efforts to incentivize clinical trial sites to shorten recruiting timelines, identifying interested participants, adequately providing informed consent, and then conducting a medical history, physical examination, laboratory, and other diagnostic studies to assess eligibility criteria is often a laborious process requiring intensive review by research staff.

Moreover, the dramatic increase in the availability of electronic health records (EHR) due to advances in information technology [62] has complicated the task of examining available data for identifying and pre-screening potential research participants. Ostensibly, the growth of health records has created both challenges and opportunities [63]. The International Classification of Disease (ICD) diagnostic codes used worldwide for clinical billing, for instance, could potentially be used to identify patients who have the condition of interest. However, diagnostic codes may also be misapplied by treating clinicians [64, 65], reflecting outdated or suspected but unconfirmed diagnoses. This inconsistency within EHRs not only complicates efforts for maintaining an accurate clinical record but also affects the ability of research staff to leverage large databases to accurately pre-screen for clinical trials. Automated methods for maintaining an accurate medical history could be a particularly useful innovation.

Given these challenges, ML techniques capable of automatically screening the EHR from prospective participants are beginning to reshape the recruitment landscape in clinical trials. These advances are predominantly driven by NLP. Though a fuller discussion of ML-driven language processing and its applications in medicine may be found beyond this paper [66], NLP is now tapping into an ability to use large amounts of "unstructured" text data, such as that used in clinical notes, whereas previous generations of ML models may have required more strictly formatted data inputs. Text sources such as radiology reports and physical examination summaries may be "featurized" in a variety

Table 2 Examples of machine learning algorithms

Logistic regression: Often used to quantify effect sizes in traditional statistics, logistic regression (Fig. 1a) may be used as an ML model by assigning each feature of a dataset to a specific parameter and then tuning these on a training set. Additional adjustments to the model design (i.e., parameter regularization) may be used to decrease the model's reliance on any singular feature and augment performance [44].

Support vector machines (SVM): SVMs (Fig. 1b) learn to set a decision boundary ("hyperplane") that maximizes the separation ("margin") between different sets of observations in the data space [45]. Decision boundaries may be adjusted to be nonlinear boundaries through specialized methods [46]. The relative simplicity of an SVM's decision often makes this model a good choice for avoiding overfitting in complex datasets (e.g., longitudinal neuroimaging data such as fMRI) [47].

K-nearest neighbors (KNN): The KNN algorithm (Fig. 1c) predicts the outcome from a set of input features from the *k* most similar points, where k is a small integer chosen by the investigator [48]. In practice, KNN is often used as a data-informed strategy for imputing missing values in a dataset. E.g., in longitudinal cohorts, KNN can be used to estimate missing variables (e.g. neurocognitive test scores [49]) by training subjects for whom full data is available

Decision trees: Decision trees (Fig. 1d) are essentially flow charts that can be used to predict outcomes based on branching logic. Given input feature values, the model undertakes a series of binary decisions to reach the proper outcome. Over the course of training, the model learns to navigate each branch point with increasing accuracy [50]. These models are best used when a high degree of interpretability is sought, but their performance may suffer relative to more complex modeling strategies

Random forests (RF): RFs (Fig. 1d) combine many randomly generated decision trees to provide an overall prediction. The overall final prediction is the result of averaging the results from individual trees in the forest through procedures such as majority voting or arithmetic averaging [51]. So-called "boosting" algorithms [52] are often used in RFs to generate successive trees that minimize the errors of earlier trees and provide improved performance. Boosted RFs are often a good benchmark for non-neural network performance, with growing deployment in academic studies [53].

Neural networks (deep learning): Neural networks (Fig. 1e) use chains of mathematical functions (or "layers") to make predictions. When many such layers are connected, the network is deep. Deep learning is a massive field that powers most of modern ML, and new "architectures" for networks are constantly emerging. However, we may note some crucial categories as follows:

- **Multilayer perceptrons (MLP):** These networks consist of an input layer of feature values that undergo further calculations in "hidden layers" before they are used to make a final prediction [54]. Such relatively simple networks may be used as standalone models or as parts of larger deep learning models
- **Convolutional neural networks (CNN):** CNNs use a series of filters that distill data patterns to their most essential properties. This is particularly useful in computer vision, where complex patterns of pixels (e.g., abscesses and organ boundaries) must be learned by building them up from simpler shapes (e.g. lines and edges) [55]. As such, these networks are best-use standards for tasks such as visual diagnosis and segmentation of critical structures in radiologic scans
- **Recurrent neural networks (RNN):** RNNs process sequential data using looped calculations that allow the network to form a "memory" of its previous processing behavior. This is highly useful for linguistic data and time series such as electrocardiograms [56] or electroencephelograms
- **Transformers:** Transformers are relatively new neural network models that differentially weight pieces of their input data using a concept known as "attention" [25]. This allows the appearance of contextual awareness in a model, such as in language tasks [58] (e.g., differing importance of different parts of a sentence) or computer vision [58] (e.g., differing importance of different areas of an image)

of formats, for instance, by scoring each document by the number of occurrences of unique words. More advanced deep learning-based methods such as large language models (i.e., BERT [67], GPT [68]) are being developed to accurately learn numerical encodings of individual words based on sentence context, thus endowing the next generation of neural networks with an ability to represent nuanced meaning in text.

NLP approaches are already being employed to derive insights from unstructured text data in clinical trials. IBM's Watson supercomputer, for instance, has been shown in recent work to improve the efficiency of patient-trial matching, increasing monthly enrollment in clinical breast cancer trials by 80% using a combination of administrative patient records and eligibility criteria from ClinicalTrials.gov [69]. Similar performance has been shown in lung cancer, as well, where Watson recently achieved 91.6% accuracy in matching eligible patients to appropriate trials [70]. Remarkably, Watson achieved such performance by matching > 7000 separate patient attributes (including histologic reports, demographics, medical/surgical history, and genomics) with > 11,000 eligibility criteria across ten phase I–III trials. With an average runtime of 15.5 s per patient, the automated approach balanced remarkable accuracy with unprecedented speed, thus hinting at the possibility of greatly reduced timelines for patient recruitment.

Automated NLP tools for study recruitment are also being used directly by patients and clinicians, as certain research groups have begun to produce tools capable of translating simple queries into computer code which can be cross-referenced with online databases of study eligibility. Researchers at Columbia University, for instance, have developed open-source tools [71] to automatically match patients with studies on ClinicalTrials.gov. Enabling non-technical usage of NLP algorithms through online search tools has the potential to streamline the tedious process of determining one's eligibility and may also democratize the usage of AI for key stakeholders. Similarly, several groups have demonstrated the viability of integrating NLP algorithms into the EHR platforms used by healthcare providers in routine care. By correlating

the information contained within a patient's medical record to databases of ongoing clinical trials, it is possible to create automated "alert systems" that flag a patient's eligibility for participation in trials of interest [72, 73].

Work in ML-based simulation methods has also suggested ways in which eligibility criteria themselves may be adjusted to streamline patient enrollment for clinical trials. A recent study by Liu and colleagues ran thousands of simulations using published eligibility criteria from a database of > 60,000 patients participating in drug trials for advanced non-small cell lung cancers [60]. In order to elucidate the influence of individual eligibility criteria on trial outcomes, the authors adapted a statistical technique developed to quantify the influence of individual features on ML model predictions [74]. By systematically identifying the importance of each criterion, they were able to identify a core set of "data-driven" conditions that increased the number of eligible patients while minimally affecting the observed hazard ratios. Work such as this carries broad importance for clinical trial research by automatically highlighting criteria that study organizers can relax conditions for patient participation. Less stringent criteria will not only help to lower barriers to study recruitment but are also likely to increase the external validity of clinical studies given that poorly designed exclusion criteria may result in systematic biases within experimental populations.

Lastly, in an age of increasing awareness of healthcare inequality, ML methods for patient recruitment may be applied to alleviate racial disparities in clinical trials. Notably, it has been estimated that nearly 90% of participants in these studies are White [75], while historical surveys of clinical trials show that they are poorly representative of women, ethnic minorities, and patients outside of relatively wealthy regions such as North America or Western Europe [76, 77]. There is little doubt that drug and medical device development poses the risk of further alienating disadvantaged patient populations when ML-based methods used to validate them in clinical trials rely on data from non-representative groups [78, 79]. The generalizability gap, however, may in part be alleviated by automated methods for improving enrollment of historically underserved groups. Zhang and colleagues, for instance, have demonstrated the usage of ML classifiers to explicitly match pregnant women and persons living with HIV to oncology trials from ClinicalTrials.gov [80]. Health systems may also use enhanced screening capacity for trial eligibility to match patients from excluded groups to ongoing studies, either by NLP methods that explicitly take into account patient identities or from the types of data-driven eligibility expansions proposed by Liu and colleagues [81, 82]. Electronic phenotyping of disease characteristics rather than demographic factors may also identify which patients are most appropriate for enrollment on the basis of their physical health, though certain clinical phenotypes (e.g., poor pulmonary function and high BMI) may retain confounding relationships with race, ethnicity, class, and gender [83]. To enhance diversity in clinical trials, a promising strategy is to use ML to identify clinical sites that may benefit from focused resources aimed at training and recruiting investigative site personnel from underrepresented minority groups. These efforts can lead to a greater representation of diverse participants in clinical trials, underscoring the importance of prioritizing such initiatives.

Learning point 2: Machine learning techniques may help improve the efficiency of clinical trials by increasing the ease of recruiting research participants.

Learning point 3: Natural language processing techniques can help identify eligibility criteria from large quantities of electronic health records and then automatically connect an individual to ongoing studies. Simulation work in this area has also shown ways in which to relax overly stringent eligibility criteria without impacting study outcomes.

Learning point 4: Natural language processing techniques can identify participants from large databases and may help alleviate racial inequities in clinical trials.

Going Beyond In-Person Trials: ML and Simulation

Given the time and expense associated with completing clinical trials, many investigators have turned their attention to alternative study designs for validating new therapies and diagnostics. With the increasing availability of large-scale health databases, novel strategies are now emerging to identify effective interventions for patients without the need to organize prospective trials. In addition, regulatory bodies are increasingly recognizing the value of such real-world evidence (RWE) as complementary to clinical trial-based evidence to support substantiation of a drug's efficacy [84]. Nevertheless, ML models are subject to the same systematic issues in data collection that plague traditional statistical analyses, such as confounding, selection bias, and inconsistent data quality [85-87]. Therefore, without carefully controlled randomization, in what ways might a new generation of predictive algorithms enable the completion of simulated clinical trials to robustly compare healthcare interventions? Could ML spur the development of a new generation of virtual or simulated trials still capable of producing trustworthy results?

Already, there is widespread interest in using external datasets to augment the statistical power of traditional clinical studies, especially in rare diseases where parallel-arm, placebo-controlled studies may be limited by the number of trial participants available [88-90], including significant support from regulatory bodies in the USA, Canada, and Europe [91]. ML technologies such as NLP may help to advance these efforts by identifying cohorts in retrospective datasets who match the eligibility criteria of patients being treated in target trials [92]. Though additional efforts are likely required

to ensure the comparability between the live and simulated study groups [93], synthetic cohorts may help to strengthen inferences in clinical studies where control groups cannot feasibly be recruited due to trial logistics for a low number of participants, such as in rare diseases [94].

Promising results are also being reported at the nexus of ML and causal inference (CI), a subfield of statistics dedicated to the identification of cause and effect in observational data [85, 86]. The fundamental challenge of CI is to quantify the difference between two separate outcomes: one that was observed (i.e., factual) and one that was not observed (i.e., counterfactual). Such a hypothetical inference may be estimated by scoring the likelihood of an individual receiving treatment (the so-called propensity score [95]) and then comparing clinical outcomes between similarly scored groups of treated and control patients [96]. Yet while such matching strategies have been shown to recapitulate the results of RCTs from observational data [97], calculating propensity scores by traditional methods may become difficult as the number of clinical variables collected from each patient becomes large [98]. Thus, ML models may also be used to derive enhanced estimates of these metrics by learning to predict treatment assignments from large quantities of data. Deep learning may even be used to provide simulated patients with propensity score matching, thus enabling the expansion of observational datasets with semisynthetic comparison groups to estimate treatment effects [99].

Lastly, a variety of research groups have now shown the capability of neural networks to learn shared patterns of characteristics (i.e., representations) between subjects receiving different forms of treatment [100]. After optimizing the identification of commonalities between patients in different treatment arms, these networks may then be used to quantify the effects of different interventions by simulating clinical outcomes in the presence or absence of a given treatment [100-102]. Such approaches essentially create "digital twins," or virtual avatars, of individual patients that may then safely be subjected to experimental therapeutics [103-105]. Still early in development, these systems may one day provide accurate, unbiased estimates of treatment effects from readily available retrospective datasets. Though time will tell, the ability to draw causal inferences by ML-driven simulations could help prioritize or modify the design of interventional RCTs by simulating the prior probability of success of an intervention without the need to even enroll a single patient.

Learning point 5: Combining machine learning with causal inference techniques can help investigators to assess cause-and-effect from observational data. This synergy can facilitate investigators in assessing the impacts of medical treatments without the need to organize large prospective studies.

Innovating Trial Design: Remote Monitoring, Digital Biomarkers, and Therapeutic Software

ML may also be used to improve the efficiency of clinical trials by alleviating many of the burdens associated with traditional, centralized study designs. In the era of COVID-19, for instance, researchers have discovered that many of the tasks previously required of patients may be completed via remote telemedicine, including the processes of obtaining informed consent [106], administering experimental drugs [107], and completing study questionnaires [72]. Given that factors such as severe illness and travel burden may contribute to patient dropout in clinical trials, remotely conducted trial visits may help investigators to retain study participants and increase the odds of a successful trial. However, when study visits are not being overseen in the clinic by research personnel, automated methods may also be able to provide quality control and ease administrative tasks.

There are myriad ways in which ML can aid remote trial administration. The US Food and Drug Administration (FDA), for instance, recently developed a mobile application (*MyStudies*) to support informed consent during the coronavirus pandemic [108]; the security of such systems may conceivably be improved by training image classification algorithms to confirm the veracity of patient signatures. Similar approaches have been adopted to confirm adherence to medication regimens in patient populations such as those experiencing mental illness or substance use disorders. As an example, AiCure, an analytics company specializing in remote clinical trial support, has employed facial recognition technology to confirm whether patients with opioid addiction are adhering to assigned medication regimens [109]. Tokyo-based Otsuka Pharmaceuticals has also piloted the usage of ingestible sensors in order to monitor the ingestion of antipsychotic drugs in patients with schizophrenia [110].

Remote monitoring of factors such as vital signs and blood chemistry could also provide early detection of adverse events in clinical trials by automatically flagging dangerous fluctuations in a participant's state of health [72]. Given the power of ML systems to detect anomalies in continuous signals [111], software programs that learn a patient's unique physiologic patterns from wearable or implantable sensors may lead the way for personalized warning systems during experimental drug trials. Additionally, ML models can learn entirely new patterns from standardly collected data, giving rise to a new generation of digital biomarkers [112, 113], to monitor treatment responses. Automated systems may learn to detect these biomarkers from a singular data source (e.g., electrocardiogram) or from combinations of multiple modalities (e.g., pulse oximetry, skin conductance, and blood glucose) to maximize the amount of information used for decision-making. In addition, physiological signals

or digital markers of real-world function, such as the use of wearable sensors to quantify mobility, may ultimately serve as clinical efficacy outcomes themselves [114, 115]. Regardless, ML may enhance the ability of the clinicians to ensure the safety of a clinical trial participant who is taking part from home and is not in the clinic.

Finally, evidence is emerging that new digital technologies may act as treatments themselves rather than simply supporting the development of traditional drugs and devices. Such "digital therapeutics" [116], including prescription video games and mobile applications, are now in the pipeline to treat conditions as diverse as ADHD, addiction, psychosis, and multiple sclerosis (MS) [117]. Though not all digital therapeutics use ML algorithms to carry out treatment, there is increasing consensus that ML technology will be required for these products to achieve future standards of precision medicine [118], and developers of these technologies are actively partnering with AI researchers to personalize and improve their delivery [119]. FDA approval and the granting of specialized "pre-certification" pathways for developers of digital therapeutics are encouraging many companies to break into this space, including both traditional pharmaceutical firms and software startups [120]. The digital revolution, with ML at its core, may bring new players to medical innovation, inevitably bringing changes to the clinical trial landscape as they seek to validate entirely novel concepts of disease therapies.

Learning point 6: Machine learning may help to alleviate obstacles to remote participation in clinical trials by enabling more effective offsite monitoring of patient well-being and adherence to medication regimens. Algorithms can help to make sense of standard data streams (e.g., vital signs) or may be trained to derive novel digital biomarkers that can provide improved prediction for outcomes of interest. Machine learning may also accelerate development of digital therapeutics, in which software itself acts as a treatment for disease.

Case Study of AI in Clinical Trials: Applications to Neurology

The great degree of variability in the presenting symptoms of neurologic disease often renders the identification of eligible patients, monitoring of progress, and evaluation of treatment endpoints in clinical trials difficult, even when performed by experienced clinicians [121]. Indeed, the complexity of neurologic disease is a likely contributor to low rates of success in clinical trials relative to other domains of medicine [122], and projected shortages in the neurologist workforce over coming decades [123, 124] threaten to exacerbate this trend. In this context, AI methodologies offer considerable benefits for clinical trials in neurology moving forward.

With respect to eligibility and recruitment, NLP offers promise across a range of clinical trials encompassing both acute and chronic conditions. In vascular neurology, for instance, NLP has been demonstrated to successfully characterize ischemic stroke from neuroradiology reports, automatically identifying TOAST [125] subtypes [126], location and acuity [127], and critical sequelae such as hemorrhagic conversion [128]. Given that shortened treatment windows after stroke onset have been shown to dramatically reduce recruitment rates in stroke trials [129], the possibility of linking AI-tagged findings to clinical trial coordinators offers a potential avenue for screening eligible patients. Moreover, enhanced electronic phenotyping is likely to improve the power of downstream data analyses, as prior work has suggested that the heterogeneous nature of stroke subtypes may contribute to mistaken conclusions from clinical trial data [130].

In neurodegenerative disorders, as well, language processing practitioners have begun to look beyond text data and are taking advantage of the potential for voice to act as an early biomarker [131] of disease that may enhance recruitment. In Alzheimer's disease (AD), the usage of voice recordings to flag likely cases of AD has been reported using neural networks [132], thus introducing the prospect of identifying potentially afflicted patients without the need for extensive neuropsychological testing [133]. Such efforts build on non-AI-based efforts to recruit patients for AD trials via analysis of vocal features gleaned from mobile applications [134]. Similar studies have been reported in Parkinson's disease (PD), where machine learning methods have been trained to differentiate PD patients from healthy controls [135, 57]. These methods will require careful planning, including informing participants that their data may result in the detection of potential clinical diagnoses. Subsequently, close integration with clinical care services to provide counseling and adequate treatment to those participants will be required of clinical trial teams, regardless of whether these individuals choose voluntarily to participate in clinical trials.

At the nexus of deep learning and epileptology, work is also being done to adjust enrollment protocols to maximize the chances of success in clinical trials. Work by Romero and Goldenholz has proposed a deep learning model that estimated the contributions of individual patients to a study's statistical power in epilepsy trials [136]. After simulating placebo and treatment arms with digitally generated cohorts, the authors demonstrated that a neural network could be trained to efficiently compute the "signal to noise ratio" offered by enrolling patients with differing seizure frequencies in randomized trials of a novel antiepileptic agent. The result of this work led to

easily interpretable "heatmaps" demonstrating which seizure parameters in newly enrolled patients might maximize the probability of detecting a treatment effect. Notably, their conclusions suggested common patterns of patient characteristics (seizure frequency and variability) that may optimize a trial's success at the time of enrollment, regardless of the outcome metric used to assess medication response [136]. Even these measures themselves may be rethought with emerging deep learning techniques: the same research group has also shown the ability of a neural network-based scoring system to discriminate drugs from placebo using 21–22% fewer patients than required with the current gold-standard metric for assessing medication response [38].

Moreover, as in other fields, ML is being used to transition from strictly centralized trial designs in neurology as well. Derivation of digital biomarkers of neurologic disease via AI-driven pattern recognition from multimodal data (e.g., wearable devices and sensors) may enable accurate monitoring of patients in neurologic diseases with fluctuating symptomology, such as PD [137, 138], AD [134, 139, 140], and various neuromuscular disorders [141]. The ability to collect such data in an automated fashion may also allow digital biomarkers to avoid many of the imprecisions brought about by basing trial endpoints on subjective behavioral and neuropsychological testing of trial participants [114, 115]. Empatica's "Embrace2" watch, for instance, is part of a growing list of FDA-approved technologies employing AI as a core feature of its design [142]. The device uses a proprietary ML classifier for seizure monitoring using data from embedded accelerometry and electrodermal activity sensors. The underlying algorithm, which was trained using video EEG labeling by board-certified neurophysiologists surveying > 5000 h of data [143] achieved a sensitivity in prospective trials > 90% for real-time detection of convulsive activity and postictal autonomic dysfunction [144], thus enabling enhanced remote monitoring of patients suffering from seizure disorders. Digital biomarkers based on ML may also help to achieve insights in trials for rare neurologic diseases such as Duchenne muscular dystrophy, where the relative precision of machine-quantified metrics derived from wearable sensors has been suggested as a means of increasing power from small sample sizes and shortening time to endpoint [114]. Additionally, remote monitoring of AI-derived digital biomarkers may elevate patient safety for those who are frail or otherwise unable to be transported directly to clinical trial sites, thus promoting healthier "aging in place" strategies [145] for elderly participants.

In the realm of digital therapeutics, ML may also soon reinvigorate trials that use such technologies as virtual reality (VR) and immersive video games to treat neurologic diseases. Already, there is extensive literature regarding the usage of digital therapeutics in neurology [146], spanning sensorimotor rehabilitation following stroke [147-152] and MS [150], chronic pain [151, 152], depression, and epilepsy management [153]. While the majority of these platforms do not utilize ML as a core feature of their design, potential avenues do exist for its integration. Certain commercial producers of VR for neuropsychiatric applications have begun to integrate AI-driven assistants (i.e., chatbots) into the design of therapeutic video games, helping users with depression to navigate cognitive reframing tasks over the course of their treatment [119]. As interactive language models based on massive "foundation" neural networks evolve (e.g., OpenAI's ChatGPT platform [154]), the usage of such technologies is slated to increase remarkably in both commercial and research applications over the coming years [105], opening avenues by which to improve the user experience of digital therapeutics in neurology and beyond.

Lastly, given sufficiently large retrospective databases, ML technologies may be trained to recapitulate individual patient outcomes across a range of neurologic conditions, and, once calibrated, they may be used to simulate treatments or forecast progression to select suitable candidates for therapeutic interventions. Neurologic disease often follows highly individualized courses influenced by individual-level and environmental factors, as well as latent disease subtypes that may be unknown at the time of trial enrollment or yet undiscovered [155, 156]. Low success rates in antiepileptic therapy [157], for example, have often been linked to the considerable variability in seizure patterns observed between individual patients. Moreover, such heterogeneity, in combination with well-known placebo effects in epilepsy trials [158, 159] has historically complicated trials of novel antiepileptics [160]. Nevertheless, recent simulation work from Goldenholz and colleagues has exemplified the ability to model approaches to recapitulate complex phenomena such as seizure cycles and clustering from large databases of self-reported seizure data [161]. The deployment of more realistic simulated datasets for longitudinal seizure trajectories may be used in ML-based strategies [136] to identify which study designs and patient characteristics are most likely to yield successful trials. In MS, as well, ML-based digital twins generated through techniques such as representation learning [162] may represent a useful clinical tool to predict disease progression and choice of treatment options given the disease's relapsing–remitting nature [163, 164]. Notably, in a study reported by the company Unlearn.AI, a neural network trained from subjects enrolled in the placebo arms of 3 MS clinical trials, was able to create a virtual cohort of digital twins that recapitulated longitudinal disease trajectories from the original patient dataset. This work raises the possibility of shortening clinical trial timelines given the ability to quickly and arbitrarily create accurately matched control groups for retrospective cohorts undergoing a variety of experimental MS treatments. The same group

has also reported statistical indistinguishability of digital twins created from retrospective MS cohorts [162], suggesting applicability of their approach across many different neurologic disease entities. In addition to simulating control groups, clinical simulations may also be employed to ensure generalizability of trial findings to populations with different demographic compositions. As an example, Chen and colleagues (using a propensity scoring method incorporating the K-nearest neighbors algorithm) recently concluded that rates of serious adverse effects reported in a phase III trial of donepezil would be much higher had the original study been composed of a majority of nonwhite participants [165]. Such conclusions, drawn with the need to organize a physical trial in a separate population, provided useful nuance regarding the drug's safety profile [166].

Technical Challenges

Despite its many promises, significant technical, pragmatic, and regulatory hurdles remain before AI technologies become a standard component of clinical trials. The inability of ML models to adequately "explain" their outputs, the potential for AI approaches to fail in prospective validation, and a regulatory environment that must adapt to rapidly evolving developments in computational science pose challenges to implementation.

Interpretability of ML models is of central importance in earning the trust of healthcare providers and clinical trial administrators, who are at the helm of high-stakes patient care. Yet, complex models such as large neural networks often produce outputs (e.g., diagnoses, simulated patients) according to internal mathematical rules that defy the causal, mechanistic explanations that are of highest importance in human reasoning [86]. ML models are often regarded as "black boxes," [167] whose usage requires leaps of faith that exceed the traditional ethical boundaries of medicine. This does not mean that frameworks for enhanced ML explainability have not begun to emerge. A particularly promising development, for instance, has been the development of "Shapley Additive Explanations" or SHAP values [74, 168]. These metrics, along with alternative explainability metrics developed for the same purpose [169-172], provide a means by which to assess the importance of individual features to a model's ultimate product. Such an approach may be used to probe ML's reliance on individual features (e.g., socioeconomic status, race) or even on individual pixels in computer vision tasks [49], thereby contextualizing model predictions in recognizable fashion. Even still, post hoc interpretation typically requires the involvement of a human subject matter expert to verify that a computer's attributions make mechanistic sense and are free of concerning biases [167]. Solutions to the interpretability gap remain, at least in part, a matter of ethical debate [173]. But from a purely technical perspective, an early solution may involve linking explainability metrics to validated clinical markers. Our group's previous work in brain MRI, for instance, has shown the ability of various neural network-derived risk scores to closely track the deposition of amyloid plaques and neurofibrillary tangles in AD patients and produce mechanistic "disease process maps" [30, 49]. We note that these results have potential applications in the noninvasive monitoring of drug response in novel trials of AD therapies. Nevertheless, adapting general explainability tools to disease-specific benchmarks defies a one-size-fits-all approach, and implementing these strategies across the full spectrum of human disease—both neurologic and non-neurologic—will require sustained efforts and interdisciplinary collaborations.

There is also the difficulty of implementing AI in clinical trial sites, which requires them to adapt their organizational infrastructure to accommodate the use of ML. At present, the vast majority of published AI models are developed as proofs of principle from retrospective datasets [174], and establishing access to these algorithms requires that clinical support staff receive adequate training in their usage, development, and access to manageable user interfaces (e.g., mobile apps, websites), and integrated into existing operational workflows such as electronic health record systems [175, 176]. Furthermore, even following the organizational and information technology realignments necessary to translate ML models to the point implementation, prospective scrutiny remains a critical factor in ensuring that they are used properly over the course of a clinical trial. Human–computer interactions often differ substantially from a model's intended usage [177], and regular audits must be performed to ensure that AI implementation is indeed facilitating a clinical trial's administration rather than hampering it. It is essential that any discordance between preclinical performance and prospective usage (particularly in models developed using synthetic or single-institution datasets) [178] be recognized in real time and that standards for early termination of clinical trials be followed in the case of serious mismatches.

Lastly, regulatory and reporting practices for AI are in flux as governing agencies adapt to a landscape of unprecedented progress. The academic community has begun to develop reporting and protocol development guidelines for clinical trials involving AI [82, 179, 180], thus contributing to a culture of accountability surrounding medical ML among researchers. Moreover, the FDA has moved to define the new category of "Software as Medical Device" (SaMD) and has outlined an updated regulatory approach via its Digital Health Innovation Action Plan [177]. As part of this shift, the agency has outlined a specific Software

Precertification Program alongside existing review pathways [177, 181] in order to facilitate streamlined approval of products employing ML in their design. Further work, however, is likely needed in order to ensure consistent quality standards in approvals such as requirements for multi-site algorithm development, dataset auditing, and prospective validation [182]. Conversely, in the EU, uniform pathways for approval of AI-based medical devices have not been developed; instead, accredited "notified bodies" in various member states are given regulatory power to issue "Conformité Éuropéen" (CE [European Conformity]) certifications prior to usage with patients, which are then mutually recognized by member states. The European Parliament, however, has passed the General Data Protection and Regulation (GPDR) law, a stringent set of guidelines that notably requires a strong degree of explainability for algorithms to be deployed in patient care [174]. The requirement to go beyond black-box models is likely to strongly impact the regulatory and innovation environment across the EU for medical AI, despite the lack of a centralized review process.

Conclusion

As medicine matures in the information age, efforts to derive actionable insights from healthcare data will advance the traditional boundaries of clinical trials. The application of machine learning technologies will require attention to data security as well as privacy and must integrate with the wealth of knowledge found in established medical practice. Responsible development in this arena has the potential to advance the pace of scientific discovery with lasting benefits for patients, clinicians, and society at large.

Funding This work was supported by grants from the Karen Toffler Charitable Trust, the American Heart Association (20SFRN35460031), and the National Institutes of Health (RF1-AG062109, R01-HL159620, R21-CA253498, and R43-DK134273), as well as a pilot award from the National Institute on Aging's Artificial Intelligence and Technology Collaboratories (AITC) for Aging Research program.

References

1. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.
2. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning. MIT press; 2018.
3. Schwartz WB. Medicine and the computer: the promise and problems of change. Use and impact of computers in clinical medicine, 1987. p. 321–35.
4. Greene JA, Lea AS. Digital futures past the long arc of big data in medicine. N Engl J Med. 2019;381:480.
5. Nash F. Differential diagnosis: an apparatus to assist the logical faculties. Lancet. 1954;263:874–5.
6. Shortliffe E. Computer-based medical consultations: MYCIN. Vol. 2. Elsevier; 2012.
7. Miller RA, McNeil MA, Challinor SM, Masarie FE Jr, Myers JD. The INTERNIST-1/quick medical REFERENCE project—status report. West J Med. 1986;145:816.
8. Blum RL. Computer-assisted design of studies using routine clinical data: analyzing the association of prednisone and cholesterol. Ann Intern Med. 1986;104:858–68.
9. Shwe MA, et al. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. Methods Inf Med. 1991;30:241–55.
10. Papik K, et al. Application of neural networks in medicine-a review. Med Sci Monit. 1998;4:538–46.
11. Crevier D. AI: the tumultuous history of the search for artificial intelligence. Basic Books, Inc.; 1993.
12. Sudlow C, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12:e1001779.
13. Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol/Współczesna Onkol. 2015;2015:68–77.
14. Pedregosa F, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
15. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2:719–31.
16. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1929–58.
17. Xie Z, et al. Artificial neural variability for deep learning: on overfitting, noise memorization, and catastrophic forgetting. Neural Comput. 2021;33:2163–92. https://doi.org/10.1162/neco_a_01403.
18. Kernbach JM, Staartjes VE. Foundations of machine learning-based clinical prediction modeling: part II-generalization and overfitting. Acta Neurochir Suppl. 2022;134:15–21. https://doi.org/10.1007/978-3-030-85292-4_3.
19. Charilaou P, Battat R. Machine learning models and over-fitting considerations. World J Gastroenterol. 2022;28:605–7. https://doi.org/10.3748/wjg.v28.i5.605.
20. Takahashi Y, et al. Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes. Transl Psychiatry. 2020;10:294. https://doi.org/10.1038/s41398-020-00957-5.
21. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning. July 2015;37:448–456.
22. Sutskever I, Martens J, Dahl G, Hinton G. Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, June 2013, Pages III-1139–III-1147.

23. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on International Conference on Machine Learning, June 2010;807–814.
24. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60:84–90.
25. Vaswani A. et al. Attention is all you need. Adv Neural Inf Proces Syst. 2017;30.
26. Tegmark M. Life 3.0: Being human in the age of artificial intelligence. Vintage; 2018.
27. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380:1347–58.
28. Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
29. Gulshan V, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402–10.
30. Qiu S, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain. 2020;143:1920–33.
31. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science; 1985.
32. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. Nat Biomed Eng. 2022; 1–7.
33. Gottesman O, et al. Guidelines for reinforcement learning in healthcare. Nat Med. 2019;25:16–8.
34. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24:1716–20.
35. Zhou X, et al. Enhancing magnetic resonance imaging-driven Alzheimer's disease classification performance using generative adversarial learning. Alzheimers Res Ther. 2021;13:1–11.
36. Goodfellow I, et al. Generative adversarial networks. Commun ACM. 2020;63:139–44.
37. Buber, E., & Banu, D. I. R. I. Performance analysis and CPU vs GPU comparison for deep learning. In 2018 6th International Conference on Control Engineering & Information Technology, CEIT. 2018;1–6.
38. Romero J, Chiang S, Goldenholz DM. Can machine learning improve randomized clinical trial analysis? Seizure. 2021;91:499–502.
39. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 1997;30:1145–59.
40. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10:e0118432.
41. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:1–28.
42. Brunet D, Vrscay ER, Wang Z. On the mathematical properties of the structural similarity index. IEEE Trans Image Process. 2011;21:1488–99.
43. Botchkarev A. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: properties and typology. arXiv preprint arXiv:1809.03006. 2018.
44. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 2002;35:352–9.
45. Boser BE, Guyon IM, Vapnik VN. In Proceedings of the fifth annual workshop on Computational learning theory. 144–52.
46. Schölkopf B. The kernel trick for distances. Adv Neural Inf Process Syst. 2000;13.
47. Pisner DA, Schnyer DM. In Machine learning. Elsevier; 2020. p. 101–21.
48. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13:21–7.
49. Qiu S, et al. Multimodal deep learning for Alzheimer's disease dementia assessment. Nat Commun. 2022;13:3404.
50. Breiman L. Classification and regression trees. Routledge; 2017.
51. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
52. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7:21.
53. Chen T, Guestrin C. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 785–94.
54. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Netw. 1989;2:359–66.
55. LeCun Y, et al. Handwritten digit recognition with a back-propagation network. Adv Neural Inf Proces Syst. 1989;2.
56. Giles CL, Kuhn GM, Williams RJ. Dynamic recurrent neural networks: theory and applications. IEEE Trans Neural Networks. 1994;5:153–6.
57. Biswas, Som. ChatGPT and the future of medical writing. Radiology 2023;223312.
58. Han K, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell. 2022;45:87–110.
59. Kelly D, Spreafico A, Siu LL. Increasing operational and scientific efficiency in clinical trials. Br J Cancer. 2020;123:1207–8.
60. Liu R, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature. 2021;592:629–33.
61. Chaudhari N, Ravi R, Gogtay NJ, Thatte UM. Recruitment and retention of the participants in clinical trials: challenges and solutions. Perspect Clin Res. 2020;11:64.
62. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13:395–405.
63. Rajkomar A, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18. https://doi.org/10.1038/s41746-018-0029-1.
64. Jetté N, Reid AY, Quan H, Hill MD, Wiebe S. How accurate is ICD coding for epilepsy? Epilepsia. 2010;51:62–9.
65. Quan H, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health Serv Res. 2008;43:1424–41.
66. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011;18:544–51.
67. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. 2018. https://doi.org/10.48550/arXiv.1810.04805.
68. OpenAI. GPT-4 techincal report. arXiv. 2023. https://doi.org/10.48550/arXiv.2303.08774.
69. Haddad T, et al. Accuracy of an artificial intelligence system for cancer clinical trial eligibility screening: retrospective pilot study. JMIR Med Inform. 2021;9:e27767.
70. Alexander M, et al. Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients. JAMIA open. 2020;3:209–15.
71. Yuan C, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc. 2019;26:294–305.
72. Kadakia KT, et al. Virtual clinical trials in oncology—overview, challenges, policy considerations, and future directions. JCO Clin Cancer Inform. 2021;4:421–5.
73. Embi PJ, Jain A, Clark J, Harris CM. In AMIA Annual Symposium Proceedings. 231 (American Medical Informatics Association).
74. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Information Proces Syst. 2017;30.

75. Knepper TC, McLeod HL. Nature Publishing Group UK London; 2018.
76. Mccarthy CR. Historical background of clinical trials involving women and minorities. Acad Med. 1994;69:695–8.
77. Thiers FA, Sinskey AJ, Berndt ER. Trends in the globalization of clinical trials. Nat Rev Drug Discovery. 2008;7:13–4.
78. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digit Med. 2020;3:99.
79. Wiens J, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25:1337–40.
80. Zhang K, Demner-Fushman D. Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations. J Am Med Inform Assoc. 2017;24:781–7.
81. Chien I, et al. In 2022 ACM conference on fairness, accountability, and transparency. 906–24.
82. Liu X, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health. 2020;2:e537–48.
83. Mosenifar Z. Population issues in clinical trials. Proc Am Thorac Soc. 2007;4:185–8.
84. Burns L, et al. Real-world evidence for regulatory decision-making: guidance from around the world. Clin Ther. 2022;44:420–37.
85. Pearl J. Causality. Cambridge university press; 2009.
86. Pearl J, Mackenzie D. The book of why: the new science of cause and effect. Basic books; 2018.
87. Larrouquere L, Giai J, Cracowski JL, Bailly S, Roustit M. Externally controlled trials: are we there yet? Clin Pharmacol Ther. 2020;108:918–9.
88. Ventz S, et al. The use of external control data for predictions and futility interim analyses in clinical trials. Neuro Oncol. 2022;24:2 47–56.
89. Ventz S, et al. Design and evaluation of an external control arm using prior clinical trials and real-world datadesign and evaluation of an external control arm. Clin Cancer Res. 2019;25:4993–5001.
90. Lingineni K, et al. Development of a model-based clinical trial simulation platform to optimize the design of clinical trials for Duchenne muscular dystrophy. CPT Pharmacometrics Syst Pharmacol. 2022;11:318–32.
91. Rahman R, et al. Leveraging external data in the design and analysis of clinical trials in neuro-oncology. Lancet Oncol. 2021;22:e456–65.
92. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183:758–64.
93. Miksad RA, Abernethy AP. Harnessing the power of real-world evidence (RWE): a checklist to ensure regulatory-grade data quality. Clin Pharmacol Ther. 2018;103:202–5.
94. Thorlund K, Dron L, Park JJ, Mills EJ. Synthetic and external controls in clinical trials–a primer for researchers. Clin Epidemiol. 2020; 457–67.
95. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
96. Farajtabar M. et al. Balance regularized neural network models for causal effect estimation. arXiv preprint arXiv:2011.11199. 2020.
97. Carrigan G, et al. Using electronic health records to derive control arms for early phase single-arm lung cancer trials: proof-of-concept in randomized controlled trials. Clin Pharmacol Ther. 2020;107:369–77.
98. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006;74:235–67.
99. Ghosh S, Boucher C, Bian J, Prosperi M. Propensity score synthetic augmentation matching using generative adversarial networks (PSSAM-GAN). Comp Methods Programs Biomed Update. 2021;1:100020.
100. Yao L. et al. Representation learning for treatment effect estimation from observational data. Adv Neural Inf Proces Syst. 2018;31.
101. Johansson F, Shalit U, Sontag D. In International conference on machine learning. 3020–29 (PMLR).
102. Shalit U, Johansson FD, Sontag D. In International Conference on Machine Learning. 3076–85 (PMLR).
103. San O. The digital twin revolution. Nature Computational Science. 2021;1:307–8.
104. Björnsson B, et al. Digital twins to personalize medicine. Genome medicine. 2020;12:1–4.
105. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28:1773–84.
106. Chhin V, et al. Leveraging mobile technology to improve efficiency of the consent-to-treatment process. JCO Clin Cancer Inform. 2017;1:1–8.
107. Upadhaya S, et al. Impact of COVID-19 on oncology clinical trials. Nat Rev Drug Discov. 2020;19:376–7.
108. Wyner Z, et al. The FDA MyStudies app: a reusable platform for distributed clinical trials and real-world evidence studies. JAMIA open. 2020;3:500–5.
109. Beaulieu T, Knight R, Nolan S, Quick O, Ti L. Artificial intelligence interventions focused on opioid use disorders: a review of the gray literature. Am J Drug Alcohol Abuse. 2021;47:26–42.
110. Waltz E. Drugs go wireless. Nat Biotechnol. 2016;34:15–9.
111. Teng M. In 2010 IEEE International Conference on Progress in Informatics and Computing. 603–08 (IEEE).
112. Au R, Kolachalama VB, Paschalidis IC. Redefining and validating digital biomarkers as fluid, dynamic multi-dimensional digital signal patterns. Front Digit Health. 2022;3:208.
113. Bent B, et al. The digital biomarker discovery pipeline: an open-source software platform for the development of digital biomarkers using mHealth and wearables data. J Clin Transl Sci. 2021;5:e19.
114. Ricotti V, et al. Wearable full-body motion tracking of activities of daily living predicts disease trajectory in Duchenne muscular dystrophy. Nat Med. 2023; 1–9.
115. Servais L, et al. Stride velocity 95th centile: insights into gaining regulatory qualification of the first wearable-derived digital endpoint for use in Duchenne muscular dystrophy trials. J Neuromuscul Dis. 2022;9:335–46.
116. Sim I. Mobile devices and health. N Engl J Med. 2019;381:956–68.
117. Kennedy C. Pear approval signals FDA readiness for digital treatments. Nat Biotechnol. 2018;36.
118. Palanica A, Docktor MJ, Lieberman M, Fossat Y. The need for artificial intelligence in digital therapeutics. Digit Biomark. 2020;4:21–5.
119. Ortolano N. Virtual reality is the latest trend in digital therapeutics. Neuroscience. 2022.
120. Food & Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). 2019.
121. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci. 2019;40:577–91.
122. Thomas DW, et al. Clinical development success rates 2006–2015. BIO Industry Analysis. 2016;1:25.
123. Burton A. How do we fix the shortage of neurologists? Lancet Neurol. 2018;17:502–3.
124. Majersik JJ, et al. A shortage of neurologists–we must act now: a report from the AAN 2019 Transforming Leaders Program. Neurology. 2021;96:1122–34.

125. Adams HP Jr, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke. 1993;24:35–41.
126. Garg R, Oh E, Naidech A, Kording K, Prabhakaran S. Automating ischemic stroke subtype classification using machine learning and natural language processing. J Stroke Cerebrovasc Dis. 2019;28:2045–51.
127. Ong CJ, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. PLoS ONE. 2020;15:e0234908.
128. Miller MI, et al. Natural language processing of radiology reports to detect complications of ischemic stroke. Neurocrit Care. 2022;37:291–302.
129. Elkins JS, Khatabi T, Fung L, Rootenberg J, Johnston SC. Recruiting subjects for acute stroke trials: a meta-analysis. Stroke. 2006;37:123–8.
130. Mandava P, Krumpelman CS, Murthy SB, Kent TA. A critical review of stroke trial analytical methodology: outcome measures, study design, and correction for imbalances. Transl Stroke Res. 2012; 833–61.
131. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for health: the use of vocal biomarkers from research to clinical practice. Digit Biomark. 2021;5:78–88.
132. Xue C, Karjadi C, Paschalidis IC, Au R, Kolachalama VB. Detection of dementia on voice recordings using deep learning: a Framingham Heart Study. Alzheimers Res Ther. 2021;13:1–15.
133. Bachman D, et al. Prevalence of dementia and probable senile dementia of the Alzheimer type in the Framingham Study. Neurology. 1992;42:115–115.
134. Gold M, et al. Digital technologies as biomarkers, clinical outcomes assessment, and recruitment tools in Alzheimer's disease clinical trials. Alzheimers Dement. 2018;4:234–42.
135. Arora S, Baghai-Ravary L, Tsanas A. Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. J Acoust Soc Am. 2019;145:2871–84.
136. Romero J, Goldenholz DM. Statistical efficiency of patient data in randomized clinical trials of epilepsy treatments. Epilepsia. 2020;61:1659–67.
137. Kassavetis P, et al. Developing a tool for remote digital assessment of Parkinson's disease. Mov Disord Clin Pract. 2016;3:59–64.
138. Espay AJ, et al. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. Mov Disord. 2019;34:657–63.
139. Dodge H, Mattek N, Austin D, Hayes T, Kaye J. In-home walking speeds and variability trajectories associated with mild cognitive impairment. Neurology. 2012;78:1946–52.
140. Kourtis LC, Regele OB, Wright JM, Jones GB. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. NPJ Digital Med. 2019;2:9.
141. Youn B-Y, et al. Digital biomarkers for neuromuscular disorders: a systematic scoping review. Diagnostics. 2021;11:1275.
142. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3:118.
143. Onorati F, et al. Multicenter clinical assessment of improved wearable multimodal convulsive seizure detectors. Epilepsia. 2017;58:1870–9.
144. Onorati F, et al. Prospective study of a multimodal convulsive seizure detection wearable system on pediatric and adult patients in the epilepsy monitoring unit. Front Neurol. 2021;12:724904.
145. Wiles JL, Leibing A, Guberman N, Reeve J, Allen RE. The meaning of "aging in place" to older people. Gerontologist. 2012; 52:357–66.
146. Abbadessa G, et al. Digital therapeutics in neurology. J Neurol. 2022;269:1209–24.
147. Choi MJ, Kim H, Nah H-W, Kang D-W. Digital therapeutics: emerging new therapy for neurologic deficits after stroke. J Stroke. 2019;21:242.
148. Cannell J, et al. The efficacy of interactive, motion capture-based rehabilitation on functional outcomes in an inpatient stroke population: a randomized controlled trial. Clin Rehabil. 2018;32:191–200.
149. Bird M, et al. "FIND technology": investigating the feasibility, efficacy and safety of controller-free interactive digital rehabilitation technology in an inpatient stroke population: study protocol for a randomized controlled trial. Trials. 2016;17:1–6.
150. Kalron A, Fonkatz I, Frid L, Baransi H, Achiron A. The effect of balance training on postural control in people with multiple sclerosis using the CAREN virtual reality system: a pilot randomized controlled trial. J Neuroeng Rehabil. 2016;13:1–10.
151. Rezaei I, Razeghi M, Ebrahimi S, Kayedi S. A novel virtual reality technique (Cervigame®) compared to conventional proprioceptive training to treat neck pain: a randomized controlled trial. J Biomed Phys Eng. 2019;9:355.
152. Austin PD, et al. The short-term effects of head-mounted virtual-reality on neuropathic pain intensity in people with spinal cord injury pain: a randomised cross-over pilot study. Spinal Cord. 2021;59:738–46.
153. Si Y, et al. Optimising epilepsy management with a smartphone application: a randomised controlled trial. Med J Aust. 2020; 212:258–62.
154. van Dis EA, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature. 2023;614:224–6.
155. Fratiglioni L, et al. Incidence of dementia and major subtypes in Europe: a collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. Neurology. 2000;54:S10-15.
156. Ferreira D, Nordberg A, Westman E. Biological subtypes of Alzheimer disease: a systematic review and meta-analysis. Neurology. 2020;94:436–48.
157. Chen Z, Brodie MJ, Liew D, Kwan P. Treatment outcomes in patients with newly diagnosed epilepsy treated with established and new antiepileptic drugs: a 30-year longitudinal cohort study. JAMA Neurol. 2018;75:279–86.
158. Rheims S, Cucherat M, Arzimanoglou A, Ryvlin P. Greater response to placebo in children than in adults: a systematic review and meta-analysis in drug-resistant partial epilepsy. PLoS Med. 2008;5:e166.
159. Zaccara G, Giovannelli F, Schmidt D. Placebo and nocebo responses in drug trials of epilepsy. Epilepsy Behav. 2015;43: 128–34.
160. Romero J, Larimer P, Chang B, Goldenholz SR, Goldenholz DM. Natural variability in seizure frequency: implications for trials and placebo. Epilepsy Res. 2020;162:106306.
161. Goldenholz DM, Westover MB. Flexible realistic simulation of seizure occurrence recapitulating statistical properties of seizure diaries. Epilepsia. 2023;64:396–405.
162. Walsh JR, et al. Generating digital twins with multiple sclerosis using probabilistic neural networks. arXiv preprint arXiv:2002.02779. 2020.
163. Voigt I, et al. Digital twins for multiple sclerosis. Front Immunol. 2021;12: 669811.
164. Denissen S, et al. Towards multimodal machine learning prediction of individual cognitive evolution in multiple sclerosis. J Pers Med. 2021;11:1349.
165. Chen Z, et al. Exploring the feasibility of using real-world data from a large clinical data research network to simulate clinical trials of Alzheimer's disease. NPJ Digit Med. 2021;4:84.

166. Wedlund L, Kvedar J. Simulated trials: in silico approach adds depth and nuance to the RCT gold-standard. NPJ Digit Med. 2021;4:121.
167. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3:e745–50.
168. Lundberg SM, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67.
169. Biecek P, Burzykowski T. Local interpretable model-agnostic explanations (LIME). Explanatory Model Analysis; Chapman and Hall/CRC: New York, NY, USA; 2021. p. 107–23.
170. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision 2017;618–626).
171. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv: 1706.03825. 2017.
172. Montavon G, Binder A, Lapuschkin S, Samek W, Müller K-R. Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning. 2019. p. 193–209.
173. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep. 2019;49:15–21.
174. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56.
175. Alhashmi SF, Alshurideh M, Al Kurdi B, Salloum SA. In Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020). 37–49 (Springer).
176. Wolff J, Pauling J, Keck A, Baumbach J. Success factors of artificial intelligence implementation in healthcare. Front Digit Health. 2021;51.
177. He J, et al. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019;25:30–6.
178. Chen RJ, Lu MY, Chen TY, Williamson DF, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5:493–7.
179. Rivera SC, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health. 2020;2:e549–60.
180. Norgeot B, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26:1320–4.
181. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health. 2021;3:e195–203.
182. Wu E, et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021;27:582–4.