Published online: 2019-08-16

AI in Health: State of the Art, Challenges, and Future Directions

Fei Wang[1], **Anita Preininger**[2]

[1] Division of Health Informatics, Department of Healthcare Policy and Research, Weill Cornell Medicine, Cornell University, NY, USA

[2] IBM Watson Health, Cambridge, MA, USA

Summary

Introduction: Artificial intelligence (AI) technologies continue to attract interest from a broad range of disciplines in recent years, including health. The increase in computer hardware and software applications in medicine, as well as digitization of health-related data together fuel progress in the development and use of AI in medicine. This progress provides new opportunities and challenges, as well as directions for the future of AI in health.

Objective: The goals of this survey are to review the current state of AI in health, along with opportunities, challenges, and practical implications. This review highlights recent developments over the past five years and directions for the future.

Methods: Publications over the past five years reporting the use of AI in health in clinical and biomedical informatics journals, as well as computer science conferences, were selected according to Google Scholar citations. Publications were then categorized into five different classes, according to the type of data analyzed.

Results: The major data types identified were multi-omics, clinical, behavioral, environmental and pharmaceutical research and development (R&D) data. The current state of AI related to each data type is described, followed by associated challenges and practical implications that have emerged over the last several years. Opportunities and future directions based on these advances are discussed.

Conclusion: Technologies have enabled the development of AI-assisted approaches to healthcare. However, there remain challenges. Work is currently underway to address multi-modal data integration, balancing quantitative algorithm performance and qualitative model interpretability, protection of model security, federated learning, and model bias.

Keywords

AI; health; deep learning; machine learning; natural language processing; federated learning

Yearb Med Inform 2019:16-26

http://dx.doi.org/10.1055/s-0039-1677908

1 Introduction

Artificial Intelligence (AI) refers to a set of technologies that allow machines and computers to simulate human intelligence. AI technologies have been developed to analyze a diverse array of health data, including patient data from multi-omic approaches, as well as clinical, behavioral, environmental, and drug data, and data encompassed in the biomedical literature.

Because of the potential to automate many tasks currently requiring human intervention, AI has attracted considerable interest from a variety of fields. AI methodologies are now commonly used to aid in computer vision, speech recognition, and natural language processing (NLP). In healthcare, the rapid development of computer hardware and software applications over recent years has facilitated digitization of health data, providing new opportunities [1] for the development of computational models and opportunities to use AI systems to extract insights from data.

AI technologies can simulate human intelligence at a variety of levels. Both machine learning (ML) and deep learning (DL) are subsets of AI. ML allows systems to learn from data at the most basic level. DL is a type of ML which uses more complex structures to build models. Conventional AI approaches (such as expert systems), according to Obemeyer and Emanuel [2], can "take general principles about medicine and apply them to new patients" in a manner similar to medical students in their first year of residency. ML abstracts rules from the data, similar to what a physician might experience during his residency [2].

One of the challenges associated with traditional ML methodologies, such as logistic regression or support vector machine (SVM) methods, is the need for intensive human effort for feature engineering. Feature engineering is the process of obtaining higher-level feature representations from raw patient features. DL approaches [1, 3] address this problem by adopting an end-to-end learning architecture, using raw patient data as an input and mapping it to outcomes through multiple layers of nonlinear processing units (i.e., neurons). This process minimizes human contributions to high-level feature engineering. However, humans are still essential for designing appropriate DL model architectures and for fine-tuning optimal model parameters. The effort to minimize the amount of human intervention required to design these architectures remains an ongoing challenge for the field.

2 Materials and Methods

This review includes works published over the past 3 to 5 years, according to the number of citations on Google Scholar. From this pool, five major types of data used in AI for health were identified. These data types include multi-omics data, clinical data, behavioral/wellness data, environmental data, as well as research and development data. The current state of AI related to each data type is discussed, followed by associated challenges and practical implications that have emerged over the last two years. Opportunities and future directions based on these data types are discussed.

3 AI for Common Biomedical Data Types

3.1 Multi-omics Data

Multi-omics data [4] refers to the biological process where different "-omics" data, such as genomics, proteomics, transcriptomics, epigenomics, and microbiomics are jointly collected and analyzed. In comparison to conventional single omics approaches, multi-omics offer a comprehensive understanding of biological processes. Separate omics data sources can often characterize the same or closely related biological processes. In ML, this is referred to as a multi-view setting [5], where each omic is regarded as a separate view. To integrate these inputs, either data-based integration or model-based integration is required.

Data-based Integration. Concatenation of the data from all of views, with or without transformation, can result in a single model. This integrative approach has been used successfully to combine data from single-nucleotide polymorphisms (SNPs) and messenger ribonucleic acid (mRNA) gene expression into a single matrix and explore the relationship between SNPs and mRNA to predict a quantitative phenotype (e.g., drug cytotoxicity) using a Bayesian integrative model [6].

Similarly, Mankoo *et al*. [7] developed an integrative approach using a multivariate Cox least absolute shrinkage and selection operator (LASSO) to predict remission rates and survival in ovarian cancer by integrating copy number alteration, methylation, microRNA (miRNA) and gene expression data. This group performed a survival analysis with a selected set of variables using Cox regression based on a variable selection via LASSO [7]. Shen *et al.* [8] proposed the iCluster framework for subtyping glioblastoma with three omics data types: copy number, mRNA expression, and DNA methylation data. The iCluster framework assumes all the omics data share a common set of latent variables during joint dimension reduction and data integration.

Model-based Integration. In this approach, a separate model based on each data view is built, followed by the aggregation of the model outputs. For example, the analysis tool for heritable and environmental network associations (ATHENA) [9-11] performed genomic analyses by integrating different omics data such as copy number alterations, methylation, miRNA and gene expression to identify associations with clinical outcomes such as ovarian cancer survival. In the integration process, base models and neural networks were first constructed based on each type of omic data, followed by integrative model building [6]. Wang *et al*. [12] proposed a network fusion approach for cancer subtyping, which begins by constructing patient similarity matrices. These matrices are based on mRNA expression, DNA methylation, and miRNA expression data. Matrix building is followed by an iterative nonlinear procedure to integrate the three base similarity matrices into a unified matrix, with the goal of identifying patient subtypes. Drăghici and Potter [13] proposed an ensemble approach to help predict drug resistance in HIV protease mutants. This approach builds a base of predictive models with structural features from an HIV protease–drug inhibitor complex and DNA sequence variants, and then performs majority voting according to the predictions of the base models.

Challenges, opportunities, and practical implications of AI in using multi-omics data. Despite the promising results that have been achieved so far, there are still many challenges to developing effective AI approaches for multi-omic data analysis.

- Because multi-omic data are highly heterogeneous, simple concatenation of raw data or model outputs from each view will miss the opportunity to explore the potential connections and relationships across entities in different views. Network-based approaches, which treat entities as nodes and their relationships as edges in the network, hold great promise for integrative analysis of multi-omic data [14]. Conventional network analysis algorithms, such as label propagation [15, 16], focus more on the edges/connections within the network. The recently proposed Graph Neural Network (GNN) [17], which considers both the node features and edge connections, would be of great interest in this context.
- Different from conventional weighted networks, edges (e.g., gene regulations and protein interactions) are usually rich contexts associated in a network constructed from multi-omic data. The incorporation of such contexts may complicate the analysis on the networks. Some typical network properties, such as edge weight non-negativity or transitivity, could be violated. Moreover, conventional network analysis assumes the network is pairwise, i.e., each edge only connects a pair of nodes in the network. However, in many scenarios we are also interested in investigating higher order interactions among different entities, for which case pairwise network analysis is not enough [18]. Therefore, there is huge potential to develop novel AI methodologies for analyzing multi-omics networks.

3.2 Clinical Data

AI technologies have also been used extensively in analyzing clinical data, including medical images, electronic health records (EHRs), and physiological signals.

3.2.1 Medical Images

Conventional ML approaches for analyzing medical images are often based on feature engineering, where features or descriptors of the medical images are extracted and then fed into the learning models for different tasks such as segmentation or classification. Due to advances that have revolutionized DL methodologies, an ever-increasing number of DL models have been incorporated into the medical image analysis pipeline. For example, Gulshan *et al.* [19] trained the Inception-V3 model [20], which is a deep learning model for natural image analysis, on a set of 128,175 renal fundus photographs for the identification of diabetic retinopathy. The authors demonstrated that, in two validation sets of 9,963 images and 1,748 images, the algorithm had 90.3% and 87.0% sensitivity, and 98.1% and 98.5% specificity, respectively. Esteva *et al.* [21] applied the same model to a set of skin images to enable discrimination between benign and malignant lesions. They designed a transfer-learning mechanism which pretrains the convolu-

tional layers of the Inception-V3 model with trained weights from ImageNet, and then retrains the final, softmax layer using a local skin image data set, fine-tuning the model parameters across all layers. Using 127,463 training images and 1,942 testing images, they demonstrated that the model can discriminate between benign and malignant lesions at a level of accuracy similar that of dermatologists. Interestingly, Kermany *et al.* [22] also adopted the same model and transfer learning strategy on two-dimensional optical coherence tomography images by freezing the parameters on the convolution layers after pretraining, without any fine tuning. With 108,312 training images and 1,000 testing images, the authors found that the model demonstrated an area under the receiving operating characteristic curve (AUC) of 99.9%. These three works demonstrate the power of end-to-end deep learning models for medical image classification through superior quantitative performance. In clinical decision support, numbers are not enough, as clinicians also need to know how the decision is made and decisions must be supported by evidence.

Recently, De Fauw *et al*. [23] proposed a novel two-stage deep learning architecture for diagnosis and patient referral (e.g., urgent, semi-urgent, routine, and observation only) of retinal disease. In the first stage, a deep segmentation network (3D Unet [24]) was developed to create a "detailed device-independent tissue segmentation map" from 3D Optical Coherence Tomography (OCT) images. Then a deep classification convolutional neural network (CNN) was constructed in the second stage to analyze the segmentation map and suggestions on diagnosis and patient referrals. After training the systems on only 14,884 scans, the approach was applied to patient triage and referral in an ophthalmology clinic. Compared with the conventional single-stage end-to-end framework, this two-stage approach derived a "device-independent segmentation of OCT scans" which serves as "intermediate representations that are readily viewable by a clinical expert" [23] and thus provides evidence for the second stage of disease diagnosis or patient referral. This facilitates the integration of the system into clinical workflows.

Challenges, opportunities, and practical implications of AI in using medical images. According to a recent report in *The Lancet*, a dermatologist may review over 200,000 images of skin lesions over decades of work, compared to mere days that it could take for a computer to analyze the same images using AI-assisted techniques [25]. ML approaches have also been used to successfully analyze raw images in cardiovascular imaging studies. By expanding the size and variety of cardiovascular imaging databases, new DL approaches can be developed, according to Heglin and colleagues [26].

Challenges remain regarding the use of AI in medical imaging. Analysis of medical images relies heavily on deep learning architectures that were designed and trained on natural images, such as the inception-V3 model discussed above. Medical images are also used to further fine-tune models. This enhances the model's ability to recognize image patterns in the training data but may not be generalizable to new image patterns. Moreover, there are few dedicated DL model architectures for medical image analysis. An associated challenge is that training a brand-new model architecture typically needs a large number of images [26], which may not be easy to obtain in medical applications.

In addition to the model challenges, there are also data challenges. For example, differences in images from patients with different ethnicities (e.g., light vs. dark skins) may introduce disparities in the model's decisions implicitly [27]. For example, if a skin lesion classification model is trained on a set of images composed of many more light skins than dark skins, it tends to perform better to classify light skins than dark ones.

3.2.2 Electronic Health Records

EHRs are systematic collections of longitudinal patient health information [28]. There are two types of information contained in patient EHRs: 1) structured information, which refers to the fields that contain data using existing lexicons, such as demographics, diagnosis, laboratory tests, medications, and procedures; and 2) unstructured information, which is typically free text documents such as clinical notes from physicians and nurses. In recent years, efforts have been devoted to developing AI methodologies for EHR analysis.

Conventional machine learning models for analyzing the structured information in EHRs are mostly vector based [29, 30], where patient records within a certain time window are collapsed into vectors composed of the summary statistics of the values of the features in different dimensions. One major limitation of this approach is that the temporality among the clinical events within EHRs is lost. To explore such temporality, Wang *et al*. [31] proposed to represent patient EHRs as longitudinal matrices with one dimension corresponding to the features and the other dimension corresponding to the time. Matrix factorization [31] or CNN type of approaches [32] were then developed to analyze such matrices. One big challenge for such matrix representation is the ultra-high sparsity. To handle such challenge, sequence modeling approaches, such as Recurrent Neural Networks (RNN) [33] have been used to analyze structured EHR data. Choi *et al*. [34] leveraged RNN to predict the onset risk of Congestive Heart Failure (CHF). To further enhance the model interpretability, they developed the REverse Time AttentIoN Model (RETAIN) [35] for modeling EHR sequences, so that the most recent clinical visits received the highest level of attention. Bekhet *et al*. [36] tested the generalizability of RETAIN on CHF onset risk prediction with a larger patient cohort. One limitation of RNN-based models is that they are not good at capturing long-term dependencies for the events in sequences. To solve this problem, Xiao *et al*. [37] leveraged TopicRNN [38], which combines RNN and global topic modeling to predict CHF patient readmission risk using EHR sequences, where each global topic corresponds to a specific distribution of the events in the EHR sequence.

Analyzing the unstructured information in EHR has been a long-standing topic in medical informatics. The conventional NLP approaches have been mostly rule-based or regular-expression-based. These methods typically need rigorous definitions of rules or regular expressions before the analysis. One challenge of these approaches is that it is impossible to enumerate all possible rules/regular expressions. In recent years, because

of the huge success of AI methods in NLP, more and more data-driven methodologies are developed for clinical NLP. For example, Kaur *et al*. [39] developed a NLP algorithm that can automatically identify patients who meet asthma predictive index (API) criteria from patient EHRs. Luo *et al*. [40] proposed to represent high-order semantic features from clinical texts as graphs and developed a subgraph-augmented nonnegative tensor factorization approach to analyze them. They also proposed segmented CNN [41] and RNN [42] to process short clinical notes and achieved state-of-the-art performance on relation classification. Filannino and Uzuner [43] performed a survey on the shared tasks for clinical NLP and identified data-driven approaches for tackling those tasks. Soysal *et al*. [44] developed a clinical language annotation, modeling, and processing (CLAMP) toolkit for customized clinical NLP applications.

Challenges, opportunities, and practical implications of AI in using EHRs. Despite promising initial results, many challenges still remain for developing AI algorithms for EHR analysis. We list some of them below.

- There are many different EHR systems all over the world. Different EHR systems may use different coding systems to encode the clinical events. The interoperability of AI algorithms across different EHR systems is critical but also challenging. There are several national/international efforts for addressing this challenge. As an example, Observational Health Data Sciences and Informatics (OHDSI, https://ohdsi.org/) is an international collaborative effort for standardizing the EHR with a common data model called Observational Medical Outcomes Partnership (OMOP). Currently it has already included 1.26 billion patient records from 17 participating countries.
- EHR data are heterogeneous, sparse, and noisy. Deriving robust AI algorithms that can reliably analyze EHR data is a challenging task. To address this challenge, interpreting or explaining how AI algorithms work is crucial, as this can provide evidences on how the algorithms make decisions [45]. Another important route is to incorporate existing medical knowledge [30] which can guide the model learning process towards the right direction.

3.2.3 Physiologic Data

Physiologic data refer to the signals from processes such as electrocardiograms (EKGs) and electroencephalograms (EEGs). These signals are usually categorized as continuous, in terms of time and value. Conventional signal processing methods usually transform those continuous-time signals into vectors through some transformations (e.g., Fourier or wavelet transform [46-48]), and then build analysis algorithms on top of these vectors. Recently, deep-learning based technologies have been used to analyze raw signals. For example, Hannun et al. [49] proposed a 34-layer CNN model to map EKG signals to a series of rhythm classes to detect heart arrhythmia. Schwab et al. [50] proposed to tackle the same problem with RNN techniques. Schirrmeister et al. [51] proposed to leverage CNN modeling to encode and visualize EEG signals. To leverage more available data, Liang et al. [52] developed a transfer learning strategy that leverages EEG data sources for seizure prediction using CNN models.

Challenges, opportunities, and practical implications of AI in using physiological data. Different from EHR, physiologic data are continuous and dense. Therefore, the analysis of physiological signals is computationally much more expensive. Preprocessing steps, such as denoising and calibration, are usually necessary before the analysis starts. Moreover, measurement errors from different devices may affect the accuracy and correctness of the analysis results. Developing approaches for modeling and reducing measurement errors is important for physiological data analysis [53].

On the other hand, the current research on analysis of physiological data typically occurs independently from analysis of other clinical data. In reality, different data may contain complementary information of the patient conditions. Therefore, performing integrative analysis of both physiological signals and other clinical data [54] would help us get a more comprehensive understanding of the patient condition, and developing effective computational approaches for such integrative analysis remains a great opportunity.

3.3 Behavioral Data

In addition to multi-omics and clinical data, behavioral data is also linked to health status. While the use of behavior data in health applications poses some specific challenges, due to the way such data is collected and housed, there are some research teams that investigate the relationship between behavior data and health.

Social Media. The use of social media, such as Facebook, Twitter, LinkedIn, and Instagram may differ according to health status. For example, Sinnenberg *et al*. [55] identified associations between Twitter posts and the risk of cardiovascular disease. From a set of 4.9 million tweets, this group found that users with cardiovascular disease can be characterized by the tone, style, and perspective of their tweets, as well as some basic demographics. Ra *et al*. [56] found "a significant association between higher frequency of modern digital media use and increase in symptoms of ADHD (attention-deficit/ hyperactivity disorder) over a 24-month period" in adolescents between the ages of 15 and 16, as compared to baseline. Researchers have examined social media analytics and mental health, and they identified markers in social media activity associated with worsening psychotic symptoms [57], schizophrenia [58], risk of suicidal ideation [59], and depression [60].

Video and Conversational Data. Use of video and conversational data has gained the attention of many, both inside and outside of fields such as healthcare. Tencent, the Chinese tech giant, claims to have developed a vision system that can spot Parkinson's Disease in 3 minutes [61]. Recently, a clinical trial involving extensive interviews between patients and trained medical staff using linguistic markers as screening tools for mild cognitive impairment (MCI) detection has shown promise [62, 63]. Tang *et al*. [64] built a conversational agent based on transcripts from these clinical trials using reinforcement learning techniques [65]. This agent was trained to maximize the diagnosis accuracy of MCI with a minimum number of conversational events, and the agent performed significantly better than supervised learning models.

Mobile Sensor Data. Many research works in recent years tried to leverage data from mobile sensors in an effort to revolutionize healthcare [66]. The insights extracted from these mobile data could be very helpful in chronic conditions such as mental health problems, chronic pain, and movement disorders. For example, Saeb *et al*. [67] studied the correlation between GPS location, phone usage data, and depressive symptom severity. Selter *et al*. [68] developed an mHealth app for self-management of chronic lower back pain. Zhan et al. [69] developed an app from mobile sensor data to quantify the Parkinson's disease severity with a machine learning approach. Turakhia and Kaiser [70] envisioned how mobile health can transform the care of atrial fibrillation. As evidence of the importance of mobile data analysis in health, the Mobile Sensor Data-to-Knowledge (MD2K) Center was chosen as one of 11 Big Data Centers of Excellence by the National Institutes of Health [71].

Challenges, opportunities, and practical Implications of AI in using behavioral data. From the above summary, we can see that behavioral data are heterogeneous. Different types of behavioral data characterize a person from different aspects, thus the integrative analysis of behavioral data can provide us a more holistic view. Insel [72] proposed the concept of digital phenotyping, which "involves collecting sensor, keyboard and voice and speech data from smartphones to measure behavior, cognition and mood." There will be many opportunities on this direction.

One challenge for analyzing behavioral data is the difficulty of obtaining the ground truth labels. For example, we can judge whether a person is likely to have depression from his/her posts on social media. However, we can only confirm the disease from the person's EHR. Therefore, linking behavioral data with clinical data can provide a unique opportunity to impact health, from both an individual and a population standpoint.

In addition to patient behavior, it is also interesting to analyze clinician behavioral data for the purpose of better quality of care delivery. Yeung *et al*. [73] proposed the concept of "bedside computer vision," which utilizes computer vision technology to analyze clinician behaviors, such as hand-hygiene compliance, captured by video recording in hospital settings This can improve the compliance of clinicians' behavior and the guidelines.

3.4 Environmental Data

Environmental factors are important in a number of diseases, including cardiovascular disease [74], chronic obstructive pulmonary disease (COPD) [75], Parkinson's Disease [76], psychiatric disorders [77], and cancer [78]. AI technologies have been used to explore environmental data to better understand disease mechanisms and improve care quality. For example, Song *et al*. [79] explored the effect of environment on hand, foot, and mouth disease through time-series analyses. Stingone *et al*. [80] studied the association between air pollution exposures and children's cognitive skills in the United States using ML models. Park *et al*. [81] leveraged advanced ML models to construct environmental risk scores and applied them to metal mixtures, oxidative, and cardiovascular disease. Hahn *et al*. [82] developed multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions.

Challenges, opportunities, and practical implications of AI in using environmental data. While the use of environmental data in AI in health holds much promise, it is not without challenges. One big challenge is to link environmental data with individual patient EHRs, given the difficulties involved in tracking the trajectories of patients and obtaining environmental information around them. Therefore, most of the studies involving environmental data are compiled at the population level. Practically, linking environmental data with other aspects of patient data may facilitate precision medicine at the patient level.

3.5 Pharmaceutical Research and Development Data

Medications play important roles in healthcare. Data collected in various stages of drug development often contain insights about disease mechanisms and treatments. AI methodologies have been adopted to extract insights from those data. Drug data are presented below according to the information source (i.e., PubChem, clinical trials, and spontaneous reports).

Chemical Compounds. PubChem [83] is a website which lists information related to small molecules and their bioactivities. Many researchers use the molecular structures contained in PubChem as a vocabulary and then adopt a footprint (zero-one) or bag-of-words representation for the analysis of specific compounds. For example, Zhang *et al*. [84, 85] used footprint-based representations to calculate drug similarities and combined them with patient or disease similarities to achieve personalized treatment recommendations. Recently, graph convolutional networks (GCN) [86] have been applied in molecular structure design and analyses, where each molecule is treated as a graph, with the atoms as graph nodes. Duvenaud *et al*. [87] designed a GCN structure to extract features (referred to as neural fingerprints) from the molecules, with good prediction capability, parsimony, and interpretability using this approach. According to Kearnes *et al*. [88], molecular graph convolutions "represent a new paradigm in ligand-based virtual screening."

Clinical Trials. Clinical trials are a key step in drug development. The participants in clinical trials are usually selected with strict inclusion and exclusion criteria. Clinical trial data provide a wealth of information for each pharmaceutical company. Recently, AI approaches have been used in clinical trial design and data mining. For example, Chekroud *et al*. [89] adopted feedforward feature selection and gradient boosting in cross-trial prediction of treatment outcomes in depression. Kohannim *et al*. [90] investigated the usage of a support vector machine to boost the power of clinical trials and reduce the clinical trial sample size.

Spontaneous Reports. The FDA Adverse Event Reporting System (FAERS) [91] collects information on adverse events related to specific drugs. For the last decade, FAERS has been the major resource for conducting pharmacovigilance research. Sakaeda *et*

al. [92] measured the performance of four concrete data mining algorithms used for predicting adverse events for specific drugs using FAERS data. These algorithms include proportional reporting ratio (PRR), reporting odds ratio (ROR), information component (IC), and empirical Bayes geometric mean (EBGM) algorithms. Tatonetti *et al*. [93] developed a signal detection algorithm for the identification of novel drug-drug interactions using FAERS. Zhang *et al*. [94] developed a label propagation algorithm to predict drug-drug interactions using drug similarity graphs obtained from side-effect profiles in FAERS. To further enhance the usability of FAERS, Banda *et al*. [95] mapped drug names and outcomes to standard vocabularies found in RxNorm and SNOMED-CT.

Challenges, opportunities, and practical implications of AI in using pharmaceutical R&D data. Despite existing promising research, challenges still exist for analyzing pharmaceutical R&D data as summarized above. We list a few of them below.

- Although graph convolution approaches have shown great promise in de-novo drug design, their interpretability remains a challenge. Specifically, in addition to more efficient discovery of novel drug molecules, understanding associated mechanisms of action is important. To achieve this goal, we should incorporate the domain knowledge from biology and chemistry into the model building process.
- One limitation of clinical trials is that they have very rigorous inclusion and exclusion criteria for patient recruitment. The goal is to eliminate the potential effect of confounding factors. However, this will also make the recruted patients "ideal" because of the rigorous recruiting constraints, and different from real world patients. Similarly, FAERS data is composed of a set of adverse drug reaction reports with limited information. To make the insights mined from clinical trial and FAERS data more practical and useful, it is crucial to link them with real world patient data from EHRs or claims. FDA has released a new strategic framework to advance the use of real-world evidence to support the development of drugs and biologics [96]. This will bring in lots of opportunities to develop AI methodologies for the integrative analysis of pharmaceutical R&D and real-world clinical data.

3.6 Biomedical Literature Data

Published reports in the biomedical literature are another important source of data for AI in health applications. AI technologies and NLP can be used to extract useful information from the literature to inform health research. Many studies focus on biomedical literature mining; for an early survey, refer to Cohen and Hersh [97]. Recently, due to the revolution of modern machine learning approaches, such as deep learning, especially in NLP, many advanced AI algorithms have been developed in biomedical literature mining and achieved state-of-the-art performance. There are two fundamental problems on literature mining: (i) named entity recognition and normalization, which is the problem of identifying interested named entities (e.g., diseases, genes, genetic variants) in the text and normalizing them (e.g., whether two different textual descriptions correspond to the same disease). For example, Leaman *et al*. [98] developed DNorm, which is a machine learning approach for disease name normalization based on pairwise learning-to-rank. The authors showed that comparing with traditional lexical normalization and matching approaches such as MetaMap [99] and Lucene [100], DNorm can achieve an improvement of 0.121 on micro-averaged F measures. Recently researchers have also shown that doing joint named entity recognition and normalization together can further boost the performance of both tasks [101, 102]; (ii) relation classification, which is the problem of identifying the relationships among named entities once they have been located in the literature. To deal with this problem, Singhal *et al.* [103] developed a rank aggregation approach to mine genotype-phenotype relationships from biomedical literatures, and they demonstrated a 28% performance improvement in terms of F1 measures on benchmarks. Peng and Lu [104] developed a multichannel dependency-based CNN approach for extracting protein-protein interactions from biomedical literature searches and achieved a 24.4% relative improvement in F1 measures over the state-of-the-art methods.

Challenges, opportunities, and practical implications of AI in using existing literature. In reality, a practical literature mining engine would involve both components we mentioned above, either explicitly or implicitly. As an example, Zhang *et al*. [105] developed a multi-view ensemble learning pipeline to integrate the textual features extracted from PubMed articles with models to classify clinically actionable genetic mutations found in specific patients. However, because both tasks are challenging, and the developed algorithms are error-prone, the error could accumulate across different stages in the pipeline and may result in bad system performance. Therefore, there is great potential on integrated end-to-end learning of the model parameters in different modules.

On the other hand, in contrast with the various biomedical data we introduced in previous sections, biomedical literature serves as the knowledge source derived from biological or clinical research. Injecting mined knowledge from such sources into the biomedical data modeling processes can make the developed models more reliable and generalizable. Tools such as PubMed Phrases [106], PubMed Labs [107], and LitVar [108] have recently been developed to facilitate research exploration of biomedical literature, which provides an unprecedented opportunity for the integration of knowledge and data driven insights from biomedical research.

4 AI in Health: Future Directions

4.1 Integrative Analysis

As Francis Collins envisioned in his vision about the precision medicine initiative [109], the next generation of scientists will "develop creative new approaches for detecting, measuring, and analyzing a wide range of biomedical information — including molecular, genomic, cellular, clinical,

behavioral, physiological, and environmental parameters." Data from different modalities can describe a health problem from different aspects, and by integrative mining of those heterogeneous data, holistic and comprehensive insights into health can be obtained.

Recent years have seen an increase in research and initiatives related to AI in health, integrating different aspects of clinical data [110], linking biorepositories with clinical data [111-113], and forging connections between pharmaceutical research and development with clinical data [84]. More importantly, combining knowledge and data is the key to developing successful AI algorithms for health. In contrast to other computer fields such as vision and speech analysis, where large data sets can be obtained, patient data is often limited and can vary widely. In addition, real-world health problems are typically complex. To help offset this problem, the expertise from clinicians and biologists is necessary to inform the model's learning process so that the model does not overfit the data.

4.2 Model Transparency

Traditional AI technologies, such as rule-based systems, are highly interpretable. Recent AI technologies, such as deep learning models, can achieve good quantitative performance, but are largely treated as black boxes. There are lots of debates recently on whether model interpretability is needed. For example, in a recent interview [114], Geoff Hinton, a pioneer in DL, argued that policymakers should not insist on ensuring people understand exactly how a given AI algorithm works, because "people can't explain how they work, for most of the things they do." Poursabzi-Sangdeh et al. [115] conducted a controlled randomized experiment to examine how important model interpretability is to users. Surprisingly, the results showed that there was no significant difference on users' trust of black-box and transparent models. Moreover, "increased transparency hampered people's ability to detect when a model has made a sizeable mistake." Holm [116] defended black-box models by drawing the analogy with human decision-making process, where decisions are largely subjective ("outcomes of their own 'deep learning'"). That's why today "neuroscience struggles with the same interpretability challenge as computer science."

According to the authors of the present article, there are certain areas where model interpretability may not be that important, especially in applications where AI algorithms have already demonstrated the capability to produce accurate results in a reliable and generalizable manner. However, this is not the case for health, at least in the current stage of the computational technology for healthcare analytics. For example, it has been shown that deep learning models can only achieve similar performance as logistic regression in hospital readmission tasks using EHRs [117] or claims [118]. Even for medical image analysis where deep learning models have achieved state-of-the-art performance, it is still difficult to justify the model generalization ability. That is, if the model works well on the medical image data set from one radiology center, it is not easy to justify it can still work well for another radiology center. Moreover, in most healthcare settings, final decision makers will still be human clinicians, and AI algorithms are just assisting them. Therefore, it is important to provide specific rationales for the propositions of those AI algorithms, to make the clinician feel more comfortable. Moreover, to enhance the clinical utility of AI algorithms, they should be integrated into regular clinical workflows [119].

On the other hand, the state-of-the-art performance of AI algorithms in many health applications are far from perfect. We should still encourage the exploration of black-box models to see if better performance can be achieved. In this case, post-hoc explanation techniques [45] would be helpful to interpret how the model works. One example of such techniques is knowledge distillation [120], which employed a student-teacher scheme to learn a simpler/interpretable model whose performance can approximate the performance of the complicated black-box model, from which the dark knowledge is "distilled out."

Another related issue about model transparency is ownership. As Shah *et al.* has envisioned in their perspective [121], there is a worrying trend towards proprietary algorithms which are opaque, and the developers are "reluctant to transparently report" model details. This may raise the potential risk of harm when these models are applied in clinical practice [122]. In this case "regulatory and professional bodies should ensure the advanced algorithms meet accepted standards of clinical benefit, just as they do for clinical therapeutics and predictive biomarkers", as Parikh *et al*. said in their discussion about predictive analytics in medicine [123].

4.3 Model Security

Conventionally we usually talk about the importance of protecting the security and privacy of health data, especially the data related to individual patients. With an increase in the number of AI models in health, we should also be aware of the potential security risk of those models. One example is adversarial attack, which refers to the process of constructing data that can confuse machine learning models and results in suboptimal or even incorrect decisions. For example, Sitawarin *et al*. [124] demonstrated that pollution on transportation signs can easily fool autonomous driving systems. Sun *et al*. [125] showed that slight modifications of lab values in a patient's EHR can completely alter the mortality prediction made by what is otherwise a well-trained predictor. Finlayson *et al*. [126] provide a more detailed discussion on the potential concerns about the "incentives for more sophisticated adversarial attacks" in healthcare. From the authors of the present article's perspective, it is important for (i) medical professionals to be aware of this potential risk; (ii) AI researchers to develop effective defense mechanisms in view of medical adversarial attacks; and (iii) policy makers to take into consideration the potential model security risk when they make new regulatory frameworks.

4.4 Federated Learning

Health data are widely distributed in and among health-related institutions, and each institution may be associated with a different set of stakeholders. In many cases, these

data are sensitive and cannot be aggregated. From a model-training perspective, it is desirable to have more and diverse data to inform model training.

Federated learning can assist with this challenge. According to Konečný *et al.* [127], "Federated Learning is a ML setting where the goal is to train a high-quality centralized model using training data distributed over a large number of clients". These clients often have unreliable and relatively slow network connections. Developing federated health AI technologies is both important and highly demanding. Lee *et al.* [128] developed a privacy-preserving federated patient similarity learning approach and evaluated it on MIMIC III data [129]. They confirmed that in a federated setting, proper homomorphic encryption of patient information can indeed preserve the quality of patient similarity measures.

In addition to clinical data, there are more and more patient-generated data nowadays. For example, these data can be continuously generated from wearable devices or mobile phones. In this case, patients could be reluctant to share their data on some public cloud to train a predictive model for their future health status. With federated learning, the model will be stored in the cloud. Each user can download the current version of the model and improve it locally with his/her data. The model changes will be summarized as a focused update which will be sent back to the cloud with encrypted communication. Then the focused updates from different users will be averaged to improve the model. During the entire process, all data will remain on local devices and no individual update is stored in the cloud. Therefore, the model will be continuously updated in a secure way.

4.5 Data Bias

All AI models need training data samples. Typically, the size of the training sample obtained from patients is not large enough to capture all variations across patients and complexities of their health problems. Frequently, the model trained from patients at one hospital does not apply to patients in another hospital. We usually refer to this challenge as the bias carried in the data, and such data bias remains one of the major challenges to AI in health. As pointed out by Khullar [130], such bias can also worsen health disparities.

One way to reduce bias is to collect large and diverse patient data sets. Examples of such efforts include the OHDSI project [131] we introduced in Section 2.2, as well as the national clinical research network PCORnet created by the Patient-Centered Outcomes Research Institute (PCORI) [132] which currently includes 13 clinical data research Networks (CDRNs) collecting longitudinal patient data from a range of health systems across the United States. These efforts serve as a foundation for collecting large-scale, diverse data sets needed for robust, generalizable AI models. Researchers can also reduce bias during the model building process [133] using methods such as counterfactual Gaussian Process which is developed to perform both risk prediction and conduct "what-if" reasoning for individualized treatment planning.

5 Conclusion

The interest, applicability, and promise of AI in health is evidenced in recent literature. This review emphasizes some of the important aspects for future consideration and research. The work underway to overcome challenges in AI in health shows promise, and this progress will facilitate the expanding role that AI is likely to continue to play in health, from both an individual and population standpoint.

Acknowledgement

Fei Wang's work is supported by NSF IIS-1750326.

References

1. Hinton G. Deep Learning-A Technology with the Potential to Transform Health Care. JAMA 2018;320(11):1101-2.
2. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 2016;375(13):1216-9.
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436-44.
4. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18(1):83.
5. Sun S. A survey of multi-view machine learning. Neural Comput Appl 2013 Dec 1;23(7-8):2031-8.
6. Fridley BL, Lund S, Jenkins GD, Wang L. A Bayesian integrative genomic model for pathway analysis of complex traits. Genet Epidemiol 2012;36(4): 352-9.
7. Mankoo PK, Shen R, Schultz N, Levine DA, Sander C. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. PLoS One 2011;6(11):e24709.
8. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, et al. Integrative subtype discovery in glioblastoma using iCluster. PloS One 2012 Apr 23;7(4):e35236.
9. Holzinger ER, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD. ATHENA: the analysis tool for heritable and environmental network associations. Bioinformatics 2014;30(5):698-705.
10. Kim S, Yeganova L, Comeau DC, Wilbur WJ, Lu Z. PubMed Phrases, an open set of coherent phrases for searching biomedical literature. Sci Data 2018;5:180104.
11. Turner SD, Dudek SM, Ritchie MD. ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci. BioData Min 2010;3(1):5.
12. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014 Mar;11(3):333.
13. Draghici S, Potter RB. Predicting HIV drug resistance with neural networks. Bioinformatics 2003;19(1):98-107.
14. Kidd BA, Readhead BP, Eden C, Parekh S, Dudley JT. Integrative network modeling approaches to personalized cancer medicine. Per Med 2015;12(3):245-57.
15. Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University; 2002.
16. Wang F, Zhang C. Label propagation through linear neighborhoods. IEEE Trans Knowl Data Eng 2008;20(1):55-67.
17. Battaglia PW, Hamrick JB, Bapst V, Sanchez-Gonzalez A, Zambaldi V, Malinowski M, et al. Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261. 2018.
18. Baytas IM, Xiao ,Wang F, Jain AK, Zhou J. Heterogeneous Hyper-Network Embedding. In: 2018 IEEE International Conference on Data Mining (ICDM); 2018. p. 875-80.
19. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402-10.
20. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 2818-26.
21. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification

of skin cancer with deep neural networks. Nature 2017;542(7639):115-8.
22. Kermany DS, Goldbaum M, Cai W, Valentim CS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 2018;172(5):1122-31.
23. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018;24(9):1342-50.
24. Cicek O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016(Oct. 17). p. 424-32.
25. Quer G, Muse ED, Nikzad N, Topol EJ, Steinhubl SR. Augmenting diagnostic vision with AI. Lancet 2017;390(10091):221.
26. Henglin M, Stein G, Hushcha PV, Snoek J, Wiltschko AB, Cheng S. Machine learning approaches in cardiovascular imaging. Circ Cardiovasc Imaging 2017;10(10):e005614.
27. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol 2018;154(11):1247-8.
28. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13(6):395-405.
29. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010:S106-13.
30. Sun J, Hu J, Luo D, Markatou M, Wang F, Edabollahi S, et al. Combining knowledge and data driven insights for identifying risk factors using electronic health records. AMIA Annu Symp Proc 2012;2012:901-10.
31. Wang F, N. Lee N, Hu J, Sun J, Ebadollahi S, Laine AF. A framework for mining signatures from event sequences and its applications in healthcare data. IEEE Trans Pattern Anal Mach Intell 2013;35(2):272-85.
32. Cheng Y, Wang F, Zhang P, Hu J. Risk prediction with electronic health records: A deep learning approach. Proceedings of the 2016 SIAM International Conference on Data Mining; 2016. p. 432-40.
33. Mikolov, T., M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur. Recurrent neural network-based language model. 11th Annual Conference of the International Speech Communication Association; 2010.
34. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2016;24(2):361-70.
35. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. An interpretable predictive model for healthcare using reverse time attention mechanism. Advances in Neural Information Processing Systems 2016:3504-12.
36. Bekhet LR, Wu Y, Wang N, Geng X, Zheng WJ, Wang F, et al. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. J Biomed Inform 2018(June 15):11-6.
37. Xiao C, Ma T, Dieng AB, Blei DM, Wang F. Readmission prediction via deep contextual embedding of clinical concepts. PLoS One 2016;13(4):e0195024.
38. Dieng AB, Wang C, Gao J, Paisley J. A recurrent neural network with long-range semantic dependency. arXiv preprint, 2016. arXiv(1611.01702).
39. Kaur H, Sohn S, Wi CI, Ryu E, Park MA, Bachman K, et al. Automated chart review utilizing natural language processing algorithm for asthma predictive index. BMC Pulm Med 2018;18(1):34.
40. Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P. Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. J Am Med Inform Assoc 2015;22(5):1009-19.
41. Li Y, Jin R, Luo Y, Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks (Seg-GCRNs). J Am Med Inform Assoc 2018;26(3):262-8.
42. Luo Y, Cheng Y, Uzuner Ö, Szolovits P, Starren J. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. J Am Med Inform Assoc 2017;25(1):93-8.
43. Filannino M, Uzuner O. Advancing the State of the Art in Clinical Natural Language Processing through Shared Tasks. Yearb Med Inform 2018;27(1):184-92.
44. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc 2017.
45. Du M, Liu N, Hu X. Techniques for interpretable machine learning. Commun ACM. In print 2019.
46. Polat K, Güneş S. Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. Appl Math Comput 2007;187(2): 1017-26.
47. Adeli H, Zhou Z, Dadmehr N. Analysis of EEG records in an epileptic patient using wavelet transform. J Neurosci Methods 2003;123(1):69-87.
48. Shaker MM. EEG waves classifier using wavelet transform and Fourier transform. International Scholarly and Scientific Research & Innovation 2007;1(3):169-74.
49. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25(1):65-9.
50. Schwab P, Scebba GC, Zhang J, Delai M, Karlen W. Beat by Beat: Classifying cardiac arrhythmias with recurrent neural networks. In: Computing in Cardiology (CinC); 2017. p. 1-4.
51. Schirrmeister RT, Springenberg JT, Fiederer LDJ, Glasstetter M, Eggensperger K, Tangermann M, et al. Deep learning with convolutional neural networks for EEG decoding and visualization. Hum Brain Mapp 2017;38(11):5391-420.
52. Liang J, Lu R, Zhang C, Wang F. Predicting seizures from electroencephalography recordings: a knowledge transfer strategy. Healthcare Informatics (ICHI) 2016 IEEE International Conference, 2016 (Oct. 4). p. 184-191.
53. Zhang K, Gong M, Ramsey J, Batmanghelich K, Spirtes P, Glymour C. Causal discovery in the presence of measurement error: Identifiability conditions. Proceedings of The Conference on Uncertainty in Artificial Intelligence (UAI); 2017.
54. Xu Y, Siddharth B, Shriprasad RD, Kevin OM, Jimeng S. RAIM: Recurrent Attentive and Intensive Model of Multimodal Patient Monitoring Data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD); 2018. p. 2565-73.
55. Sinnenberg L, DiSilvestro CL, Mancheno C, Dailey K, Tufts C, Buttenheim AM, et al. Twitter as a Potential Data Source for Cardiovascular Disease Research. JAMA Cardiol 2016;1(9):1032-6.
56. Ra CK, Cho J, Stone MD, De La Cerda J, Goldenson NI, Moroney E, et al. Association of Digital Media Use With Subsequent Symptoms of Attention-Deficit/Hyperactivity Disorder Among Adolescents. JAMA 2018;320(3):255-63.
57. Birnbaum M, Rizvi A, De Choudhury M, Ernala S, Cecchi G, Kane J. O9.2. Identifying psychotic symptoms and predicting relapse through social media. Schizophr Bull, 2018;44 (Suppl. 1): S100.
58. Birnbaum ML, Ernala SK, Rizvi AF, De Choudhury M, Kane JM. A collaborative approach to identifying social media markers of schizophreniz by employing machine learning and clinical appraisals. J Med Internet Res 2017(19):8.
59. De Choudhury M, Kiciman E. The language of social support in social media and its effect on suicidal ideation risk. Proc Int AAAI Conf Weblogs Soc Media 2017 May;2017:32-41.
60. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. In: Seventh international AAAI conference on weblogs and social media (ICWSM) 2013;8(13):128-37.
61. Forbes. Tencent Aims to Train AI to Spot Parkinson's in 3 Minutes, in https://www.forbes.com/sites/samshead/2018/10/08/tencent-aims-to-train-ai-to-spot-parkinsons-in-3-minutes/#6db6d3f06f36; 2018.
62. Dodge HH, Zhu J, Mattek NC, Bowman M, Ybarra O, Wild KV, et al. Web-enabled Conversational Interactions as a Means to Improve Cognitive Functions: Results of a 6-Week Randomized Controlled Trial. Alzheimers Dement (N Y) 2015;1(1):1-12.
63. Asgari M, Kaye J, Dodge H. Predicting mild cognitive impairment from spontaneous spoken utterances. Alzheimers Dement (N Y) 2017;3(2):219-28.
64. Tang F, Lin K, Uchendu I, Dodge HH, Zhou J. Improving mild cognitive impairment prediction via reinforcement learning and dialogue simulation. arXiv preprint ar Xiv, 2018. 1802.06428.
65. Sutton RS, Barto AG. Introduction to reinforcement learning. Vol. Mar. 1, 1998. Cambridge: MIT Press; 1998.
66. Kumar S, Nilsen WJ, Abernethy A, Atienza A, Patrick K, Pavel M, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013;45(2):228-36.
67. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. J Med Internet Res 2015;17(7):e175.
68. Selter A, Tsangouri C, Ali SB, Freed D, Vatchinsky A, Kizer J, et al. An mHealth App for Self-Management of Chronic Lower Back Pain (Limbr): Pilot Study. JMIR Mhealth Uhealth 2018;6(9):e179.

69. Zhan A, Mohan S, Tarolli C, Schneider RB, Adams JL, Sharma S, et al. Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score. JAMA Neurol 2018;75(7):876-80.
70. Turakhia MP, Kaiser DW. Transforming the care of atrial fibrillation with mobile health. J Interv Card Electrophysiol 2018;47(1):45-50.
71. Kumar S, Abowd GD, Abraham WT, al'Absi M, Beck JG, Chau DH, et al. Center of excellence for mobile sensor data-to-knowledge (MD2K). J Am Med Inform Assoc 2015;22(6):1137-42.
72. Insel TR. Digital Phenotyping: Technology for a New Science of Behavior. JAMA 2017;318(13):1215-6.
73. Yeung S, Downing NL, Li FF, Milstein A. Bedside Computer Vision - Moving Artificial Intelligence from Driver Assistance to Patient Safety. N Engl J Med 2018;378(14):1271-3.
74. Cosselman KE, Navas-Acien A, Kaufman JD. Environmental factors in cardiovascular disease. Nat Rev Cardiol 2015;12(11):627-42.
75. MacNee W, Donaldson K. Exacerbations of COPD: environmental mechanisms. Chest 2000;117(5 Suppl 2):390S-7S.
76. Tanner CM. The role of environmental toxins in the etiology of Parkinson's disease. Trends Neurosci 1989;12(2):49-54.
77. Guloksuz S, van Os J, Rutten BPF. The Exposome Paradigm and the Complexities of Environmental Research in Psychiatry. JAMA Psychiatry 2018;75(10):985-6.
78. Boffetta P, Nyberg F. Contribution of environmental factors to cancer risk. Br Med Bull 2003;68:71-94.
79. Song Y, Wang F, Wang B, Tao S, Zhang H, Liu S, et al. Time series analyses of hand, foot and mouth disease integrating weather variables. PLoS One 2015;10(3):e0117296.
80. Stingone JA, Pandey OP, Claudio L, Pandey G. Using machine learning to identify air pollution exposure profiles associated with early cognitie skills among us children. Environmental Pollution 2017;230:730-40.
81. Park SK, Zhao Z, Mukherjee B. Construction of environmental risk score beyond standard linear models using machine learning methods: Application to metal mixtures, oxidative stress and cardiovascular disease in NHANES. Environ Health 2017;16(1):102-19.
82. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 2003;19(3):376-82.
83. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 2009;37(Web Server issue):W623-33.
84. Zhang P, Wang F, Hu J, Sorrentino R. Towards personalized medicine: leveraging patient similarity and drug similarity analytics. AMIA Jt Summits Transl Sci Proc 2014:132-6.
85. Zhang P, Wang F, Hu J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. AMIA Annu Symp Proc 2014:1258-67.
86. Bronstein MM, Bruna J, Le Cun Y, Szlam A, Vandergheynst P. Geometric deep learning: Going beyond Euclidean data. IEEE Signal Process Mag 2017;34(4):18-42.
87. Duvenaud DK, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, et al. Conventional networks on graphs for learning molecular fingerprints. Adv Neural Inf Process Syst 2015:2224-32.
88. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des 2016;30(8):595-608.
89. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 2016;3(3):243-50.
90. Kohannim O, Hua X, Hibar DP, Lee S, Chou YY, Toga AW, et al; Alzheimer's Disease Neuroimaging Initiative. Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiol Aging 2010;31(8):1429-42.
91. FDA, The FDA Adverse Event Reporting System. https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/Surveillance/AdverseDrugEffects/default.htm.
92. Sakaeda T, Tamon A, Kadoyama K, Okuno Y. Data mining of the public version of the FDA Adverse Event Reporting System. Int J Med Sci 2013;10(7):796-803.
93. Tatonetti NP, Fernald GH, Altman RB. A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. J Am Med Inform Assoc 2013;19(1):79-85.
94. Zhang P, Wang F, Hu J,Sorrentino R. Label Propagation Prediction of Drug-Drug Interactions Based on Clinical Side Effects. Sci Rep 2015;5:12339.
95. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data 2016;3:160026.
96. US Food and Drug Administration. Framework for FDA's Real-World Evidence Program 2018; 2019.
97. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform 2005;6(1):57-71.
98. Leaman R, RDoğan RI, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 2013;29(22):2909-17.
99. Aronson AR. Metamap: Mapping text to the UMLS metathesaurus. Bethesda, MD: NLM, NIH, DHHS; 2006. p. 1-26.
100. McCandless M, Hatcher E, Gospodnetic O. Lucene in action: covers Apache Lucene 3.0. Manning Publications Co.; 2010.
101. Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics 2016;32(18):2839-46.
102. Zhao S, Wang F, Zhao S, Liu T. A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI); 2019.
103. Singhal A, Simmons M, Lu Z. Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. PLoS Comput Biol 2016;12(11): e1005017.
104. Peng Y, Z Lu. Deep learning for extracting protein-protein interactions from biomedical literature. Proceedings of the ACL BioNLP 2017 workshop; 2017. p. 29–38.
105. Zhang X, Chen D, Zhu Y, Che C, Su C, Zhao S, et al. Multi-view ensemble classification for clinically actionable genetic mutations. The NIPS'17 Competition: Building intelligent systems; 2018. p. 79-99.
106. Kim S, Yeganova L, Comeau DC, Wilbur WJ, Lu Z. PubMed Phrases, an open set of coherent phrases for searching biomedical literature. Sci Data 2018;5:180104.
107. Fiorini N, Canese K, Bryzgunov R, Radetska I, Gindulyte A, Latterner M, et al. PubMed Labs: an experimental system for improving biomedical literature search. Database (Oxford) 2018;2018.
108. Allot A, Peng Y, Wei CH, Lee K, Phan L, Lu Z. LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic Acids Res 2018;46(W1):W530-W536.
109. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372(9):793-5.
110. Zhang X, Chou J, Wang F. Integrative analysis of patient health records and neuroimages via memory-based graph convolutional network. In: Proceedings of IEEE International Conference on Data Mining (ICDM); 2018. p. 767-76.
111. McCarty, C.A., R. L. Chisholm, C. G. Chute, I. J. Kullo, G. P. Jarvik, E. B. Larson, R. Li, et al., The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics, 2011. 4: p. 13.
112. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013;15(10):761-71.
113. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med 2013;7(311):311ra174.
114. Simonite T. Google's AI Guru Wants Computers to Think More Like Brains. Wired Business. 12/12/2018. https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/.
115. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan J, Wallach H. Manipulating and measuring model interpretability. arXiv preprint arXiv:1802.07810;2018.
116. Holm EA. In Defense of the Black Box. Science 2019;364(6435):26-7.
117. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine 2018;1(1):18.
118. Min X, Yu B, Wang F. Predictive Modeling of the Hospital Readmission Risk from Patients' Claims Data Using Machine Learning: A Case Study on COPD. Sci Rep 2019;9(1):2362.
119. Wang F, Casalino LP, Khullar D. Deep Learning in Medicine—Promise, Progress, and Challenges. JAMA Intern Med 2019 Mar 1;179(3):293-4.
120. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531; 2015.
121. Shah N.D, Steyerberg EW, Kent DM. Big data

and predictive analytics: recalibrating expectations. JAMA 2018;320(1):27-8.
122. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56.
123. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. Science 2019;363(6429):810-2.
124. Sitawarin C, Bhagoji AN, Mosenia A, Chiang M, Mittal P. Darts: Deceiving autonomous cars with toxic signs. arXiv preprint arXiv:1802.06430; 2018.
125. Sun M, Tang F, Yi J, Wang F, Zhou J. Identify susceptible locations in medical records via adversarial attacks on deep predictive models. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD); 2018. p. 793-801.
126. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363(6433):1287-9.
127. Konecny J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: Strategies for improving communication efficiency. ArXiv preprint. arXiv(1610.05492); 2016: p.18.
128. Lee J, Sun J, Wang F, WangS, JunCH, Jiang X. Privacy-Preserving Patient Similarity Learning in a Federated Environment: Development and Analysis. JMIR Med Inform 2018;6(2):e20.
129. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.
130. Khullar D. A.I. could worsen health disparities. The New York Times. Jan. 31, 2019. https://www.nytimes.com/2019/01/31/opinion/ai-bias-healthcare.html.
131. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-8.
132. PCORI: Patient-Centered Outcomes Research Institute. https://www.pcori.org/.
133. Schulam P, Saria S. Reliable decision support using counterfactual models. Adv Neural Inf Process Syst 2017:1697-708.

Correspondence to:
Fei Wang, PhD, Associate Professor
Division of Health Informatics
Department of Healthcare Policy and Research
Weill Cornell Medicine
Cornell University
425 East 61 Street
New York. NY 10065, USA
E-mail: few2001@med.cornell.edu