Natural Language Processing for EHR-Based Computational Phenotyping

Zexian Zeng, Yu Deng, Xiaoyu Li, Tristan Naumann, Yuan Luo

Abstract— This article reviews recent advances in applying natural language processing (NLP) to Electronic Health Records (EHRs) for computational phenotyping. NLP-based computational phenotyping has numerous applications including diagnosis categorization, novel phenotype discovery, clinical trial screening, pharmacogenomics, drug-drug interaction (DDI) and adverse drug event (ADE) detection, as well as genome-wide and phenome-wide association studies. Significant progress has been made in algorithm development and resource construction for computational phenotyping. Among the surveyed methods, well-designed keyword search and rule-based systems often achieve good performance. However, the construction of keyword and rule lists requires significant manual effort, which is difficult to scale. Supervised machine learning models have been favored because they are capable of acquiring both classification patterns and structures from data. Recently, deep learning and unsupervised learning have received growing attention, with the former favored for its performance and the latter for its ability to find novel phenotypes. Integrating heterogeneous data sources have become increasingly important and have shown promise in improving model performance. Often better performance is achieved by combining multiple modalities of information. Despite these many advances, challenges and opportunities remain for NLP-based computational phenotyping, including better model interpretability and generalizability, and proper characterization of feature relations in clinical narratives.

Index Terms—Electronic Health Records, Natural Language Processing, Computational Phenotyping, Machine Learning

1 INTRODUCTION

A phenotype is an expression of the characteristics that result from genotype variations and an organism's interactions with its environment. A phenotype may consist of physical appearances (e.g., height, weight, BMI), biochemical processes, or behaviors [1]. In the medical domain, phenotypes are often summarized by experts on the basis of clinical observations. Nationwide adoption of Electronic Health Records (EHRs) has given rise to a large amount of digital health data, which can be used for secondary analysis [2]. Typical EHRs include structured data such as diagnosis codes, vitals and physiologic measurements, as well as unstructured clinical narratives such as progress notes and discharge summaries. Computational phenotyping aims to automatically mine or predict clinically significant, or scientifically meaningful, phenotypes from structured EHR data, unstructured clinical narratives, or their combination.

As summarized in a 2013 review by Shivade et al. [3], early computational phenotyping studies were often formulated as supervised learning problems wherein a predefined phenotype is provided, and the task is to construct a patient cohort matching the definition's criteria. Many of these studies relied heavily on structured and coded patient data; for example, using encodings such as International Classification of Disease, 9th Revision (ICD-9) [4], its successor the 10th Revision (ICD-10) [5], Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [6], RxNorm [7], and Logical Observation Identifiers Names and Codes (LOINC) [8]. On the other hand, the use of natural language processing (NLP) for EHR-based computational phenotyping has been limited to term and keyword extraction [3].

Structured data typically capture patients' demographic information, lab values, medications, diagnoses, and encounters [9]. Although readily available and easily accessible, studies have concluded that structured data alone are not sufficient to accurately infer phenotypes [10, 11]. For example, ICD-9 codes are mainly recorded for administrative purposes and are influenced by billing requirements and avoidance of liability [12, 13]. Consequently, these codes do not always accurately reflect a patient's underlying physiology. Furthermore, not all patient information is well documented in structured data, such as clinicians' observations and insights [14]. As a result, using structured data alone for phenotype identification often results in low performance [11]. The limitations associated with structured data for computational phenotyping have encouraged the use of clinical narratives, which typically include clinicians' notes, observations, referring letters, specialists' reports, discharge summaries, and a record of communications between doctors and patients [15]. Unstructured clinical narratives may summarize patients' medical history, diagnoses, medications, immunizations, allergies, radiology images, and laboratory test results, in the forms of progress notes, discharge reports etc. [16].

Structured and unstructured EHR data are often stored in vendor applications or at a healthcare enterprise data warehouse. Typical

- *Zexian Zeng is with the Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611. E-mail: zexian.zeng@northwestern.edu.*
- *Yu Deng is with the Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611. E-mail: yu.deng@northwestern.edu.*
- *Xiaoyu Li is with the Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA 02115. E-mail: xil288@mail.harvard.edu.*
- *Tristan Naumann is with Computer Science and Artificial Intelligence Lab, Massachusetts Institue of Technology, Cambridge, MA 02139. E-mail: tjn@mit.edu.*
- *Yuan Luo is with the Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611. E-mail: yuan.luo@northwestern.edu.*

EHR data are usually managed by a local institution's technicians and are accessible to trained personnel or researchers. Institutional Review Boards at local institutions typically grant access to certain patient cohorts and certain parts of EHRs. Database queries can then be written and executed to retrieve desired structured and unstructured EHR data. In addition to hospital-collected data stored in EHRs, research data are increasingly available, including public databases such as PubMed [17], Textpresso [18], Human Protein Interaction Database (HPID) [19], and MeInfoText [20]. With growing amount of available data, efficient identification of relevant documents is essential to the research community. Information retrieval systems have been developed to identify text corresponding to certain topics or areas from EHR data across multiple fields. CoPub Mapper [21] ranks co-occurrence associations between genes and biological terms from PubMed. iHOP [22] links interacting proteins to their corresponding databases and uses co-occurrence information to build a graphical interaction network. We refer the reader to the following reviews for more details: [23] is a survey for biomedical text mining in cancer research, [24] is a survey for biomedical text mining, and [25] is a survey for web mining.

While the prevalence of EHR data presents an opportunity for improved computational phenotyping, extracting information from clinical narratives for accurate phenotyping requires both semantic and syntactic structures in the narrative to be captured [26]. Scaling such tasks to large cohort studies is laborious, time-consuming, and typically requires extensive data collection and annotation.

Recently, NLP methods for EHR-based computational phenotyping have seen extensive development, extending beyond basic term and keyword extraction. One focus of recent studies is formulating computational phenotyping as an unsupervised learning problem to automatically discover unknown phenotypes. The construction of richer features such as relations between medical concepts enables greater expressive power when encoding patient status, compared to terms and keywords. More advanced machine learning methods, such as deep learning, have also been increasingly adopted to learn the underlying patient representation.

This article reviews the literature on NLP methods for EHR-based computational phenotyping, emphasizing recent developments. We first describe several applications of computational phenotyping. We then summarize the state-of-the-art NLP methods for computational phenotyping and compare their advantages and disadvantages. We also describe the combinations of data modalities, feature learning, and relation extraction that have been used to aid computational phenotyping. Finally, we discuss challenges and opportunities to NLP methods for computational phenotyping and highlight a few promising future directions.

2 APPLICATIONS OF EHR-BASED COMPUTATIONAL PHENOTYPING

Computational phenotyping has facilitated biomedical and clinical research across many applications, including patient diagnosis categorization, novel phenotype discovery, clinical trial screening, pharmacogenomics, drug-drug interaction (DDI) and adverse drug event (ADE) detection, and downstream genomics studies.

2.1 Diagnosis Categorization

One of the most important applications of computational phenotyping is diagnosis categorization, which enables the automated and efficient identification of patient cohorts for secondary analysis [15, 27-31]. A wide range of diseases has been investigated in the past, including suspected tuberculosis (TB) [32, 33], colorectal cancer [34], rheumatoid arthritis [35], diabetes [36], heart failure [37, 38], neuropsychiatric disorders [39], etc. These applications have extended from disease identification to disease subtyping such as lung cancer stage evaluation [40], or subsequent event detection such as breast cancer recurrence detection [41] and cancer metastases detection [42].

2.2 Novel Phenotype Discovery

Computational phenotyping has been applied to discover novel phenotypes and sub-phenotypes. Traditionally, a clinical phenotype is classified into a particular category if it meets a set of criteria developed by domain experts [43]. Instead, semi-supervised or unsupervised methods can detect traits based on intrinsic data patterns with moderate or minimal expert guidance, which may promote the discovery of novel phenotypes or sub-phenotypes. For example, in a study by Marlin et al. [44], a diagonal covariance Gaussian mixture model was applied on physiological time series data for patient clustering. They discovered distinct, recognizable physiological patterns and they concluded that interpretations of these patterns could offer prognostic significance. Doshi-Velez et al. [45] applied hierarchical clustering to define subgroups with distinct courses among autism spectrum disorders. They applied ICD-9 codes to construct time series features. In the study, they identified four subgroups among 4934 patients; one subgroup was characterized by seizures; one subgroup was characterized by multisystem disorders including gastrointestinal disorders, auditory disorders, and infections; one subgroup was characterized by psychiatric disorders; one subgroup could not be further resolved. In a study by Ho et al. [46], they applied tensor factorization [47, 48] on medication orders to generate phenotypes without supervision. In a case study searching for 50 phenotypes in heart failure, they achieved better performance than principal component analysis (PCA) with respect to area under curve (AUC) score and model stability. Further interpretations of these novel phenotypes have potential to offer us useful clinical information. Shah et al. [49] clustered patients with preserved ejection fraction into three novel subgroups, which offers meaningful insight into clinical characteristics, cardiac structures, and outcomes.

2.3 Clinical Trial Screening

Leveraging EHR data can benefit clinical trial recruitment [50]. In recent years, echoing the rising availability of EHR data and the increased volume of clinical trial recruitments, computational phenotyping for clinical trial screening has become an active area. Multiple systems have been designed for this purpose [51-54]. Electronic screening can improve efficiency in clinical trial recruitment, and automated querying over trials can support clinical knowledge curation [55]. A typical computational phenotyping system for clinical trial eligibility identifies patients whose profiles—extracted from structured data and narratives—matched the trial criteria in order to reduce the pool of candidates for further staff screening.

2.4 Pharmacogenomics

Pharmacogenomics aims to investigate the interaction between genes, gene products, and therapeutic substances. Much of this knowledge exists in scientific literature and curated databases. Computational phenotyping applications have been developed to mine pharmacogenomics knowledge [56-59]. These phenotyping tools automatically scan, retrieve, and summarize the literature for meaningful phenotypes. Recent studies have adopted semantic and syntactic analyses as well as statistical machine learning tools to mine targeted pharmacogenomics relations from scientific literature and clinical records [58].

2.5 DDIs and ADEs

Drug-drug interactions (DDIs) happen when one drug affects the activity of another drug that has been simultaneously administered. Adverse drug events (ADEs) refer to unexpected injuries caused by administering medication. Detecting DDIs and ADEs can guide the process of drug development and drug administration. The impact of these negative outcomes has triggered huge efforts from industry and the scientific community to develop models exploring the relationships between drugs and biochemical pathways in order to enable the discovery of DDIs [60, 61] and ADEs [26, 62-64].

2.6 GWAS and PheWAS

Cohorts obtained by computational phenotyping have benefited downstream genomic studies [65], using techniques such as Genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS). In GWAS, researchers link genomic information from DNA biorepositories to EHR data to detect associations between phenotypes and genes. In such studies, case-control cohorts can be generated without labor intensive annotation, which is especially important for rare variant studies where a large number of patients need to be screened. Much research [66-69] has explored EHR phenotyping algorithms to facilitate GWAS. We refer the reader to reviews by Bush et al. [70] and Wei et al. [65] for more details. PheWAS studies analyze a wide range of phenotypes affected by a specific genetic variant. Denny et al. [71] applied computational phenotyping on EHR to automatically detect 776 different disease populations and their matched controls. Statistical tests were then carried out to determine associations between single nucleotide polymorphisms and multiple disease phenotypes. Additional studies have established the efficiency of EHR-based PheWAS to detect genetic association [72-74]. Compared to traditional genomic research, computational phenotyping has driven discovery of variant-disease associations and has facilitated the completion of genomic research in a timely and lower cost manner [66].

3 METHODS FOR NLP-BASED COMPUTATIONAL PHENOTYPING

NLP methods for computational phenotyping algorithms exhibit a wide range of complexities. Early stage systems were often based on keyword search or customized rules. Later, supervised statistical machine learning algorithms were applied extensively to computational phenotyping. More recently, unsupervised learning has resulted in effective patient representation learning and discovery of novel phenotypes. This section reviews NLP methods for EHR-based computational phenotyping, starting with three major categories: 1) keyword search or rule-based systems, 2) supervised learning systems, and 3) unsupervised systems. We then identify current trends and active directions of development. For convenience, we summarize the characteristics of studies reviewed in this section in

Table 1. The studies are characterized regarding the methods used to generate features, the methods or tools used for classifying the assertions (e.g., negations) of the features, the named entity recognition methods used to identify the concepts in the narratives, and the data sources used for modeling training.

3.1 Keyword Search and Rule-based System

Keyword search is one of the algorithms with the least model complexity for computational phenotyping. It looks for keywords, derivations of those keywords, or a combination of keywords to extract phenotypes [75]. For example, “pneumonia in the right lower lobe” is a derivation of the key phrase “consolidation in the left lower lobe” in Fiszman et al. [75]. These keywords correspond to medications, diseases, or symptoms; and, in practice, they are often identified using regular expressions. In early work, large tables of keywords were generated. Meystre et al. [76] manually built a keyword table using 80 selected concepts with related sub-concepts. They retrieved 6,928 phrases corresponding to the 80 concepts from the Unified Medical Language System (UMLS) Metathesaurus *MRCONSO* table [77]. After filtering, they still had 4,570 keywords remaining. Based on these keywords for classification, they achieved a precision of 75% and a recall of 89%. Wagholikar et al. [78] developed a keyword search system for limb abnormality identification using free-text radiology reports. Even though the reports have an average length of only 52 words, they achieved an F-measure of 80% and an accuracy of 80%. Despite their success, problems caused by the unstructured, noisy nature of the narrative text (e.g., grammatical ambiguity, synonyms, term abbreviation, misspelling, or negation of concepts) remain bottlenecks in keyword search. In general, keyword search is more susceptible to low accuracy due to simplicity of features. To improve model performance, supplementary rules (or other more sophisticated criteria) have been added to keyword search.

Rule-based systems are among the most frequently used computational phenotyping methods. In a review by Shivade et al. [3], 24 out of 97 computational phenotyping related articles have described rule-based systems. In a typical rule-based system, criteria need to be pre-defined by domain experts. For example, Wiley et al. [79] developed a rule-based system for stain-induced myotoxicity detection. They manually annotated 300 individuals’ allergy listings and pre-defined a set of keywords. Then they developed a set of rules to detect contextual mentions around the keywords. In this study, they achieved a positive predictive value (PPV) score of 86% and a negative predictive value (NPV) score of 91%. Ware et al. [80] developed a list of concepts together with a list of secondary concepts that appear in the same sentence. The secondary concepts were mainly medications. After defining the concepts, they developed a set of rules for phenotype identification. This framework achieved an overall kappa score of 92% with the original annotations. Nguyen et al. [40] implemented an NLP tool, called the General Architecture for Text Engineering (GATE), to extract UMLS concepts and mapped them to SNOMED CT concepts. These SNOMED CT concepts were utilized to predict the stage of lung cancer using defined rules based on staging guidance. They achieved accuracies of 72%, 78%, and 94% for T, N, and M staging, respectively.

Xu et al. [34] implemented a heuristic rule-based approach for colorectal cancer assertion. The system used MedLEE [81] to detect colorectal cancer-related concepts. It then applied defined rules to search for concept contexts. The system achieved an F-measure of 99.6% for document level concept identification. Li et al. [82] developed a rule-based system to detect adverse drug events and medical errors using patients' clinical narratives, medications, and lab results. They compared the model's performance to a trigger tool [83], and they achieved 100% agreement. The triggers in the trigger tool are a combination of keywords that signal an underlying event of interest. Haerian et al. [84] defined rules to extract concepts from discharge summaries on top of the ICD-9 code. The use of concepts increased the model's PPV score from 55% to 97%. Sauer et al. [85] developed a set of rules to identify bronchodilator responsiveness from pulmonary function test reports, and they achieved an F-measure of 98%.

Rule-based systems often need many complex attribute-specific rules, which may be too rigid to account for the diversity of the language expression. As a result, rule-based systems may exhibit have high precision, but low recall. In fact, as will be detailed in the next subsections, more recent systems opted to use statistical machine learning algorithms to replace or complement rules.

Developing rules is laborious, time-consuming and requires expert knowledge. Despite these disadvantages, rule-based systems remain one of the most popular computational phenotyping methods in the field due to their straightforward construction, easy implementation, and high accuracy [30].

3.2 Supervised Statistical Machine Learning Algorithms

To improve upon accuracy and scalability while decreasing domain expert involvement, statistical machine learning methods have been adopted for computational phenotyping. These methods usually have the advantage that in addition to classifying phenotypes, they often provide the probability or confidence of that classification. In general, statistical machine learning methods are categorized as supervised, semi-supervised, or unsupervised. Common to all methods, each subject is represented as a vector consisting of features. In supervised learning, each sample in a training dataset is labeled. Algorithms predict the labels for an unknown or test dataset after learning from the training dataset. In contrast, unsupervised learning identifies patterns without labeling. It automatically clusters samples with similar patterns into groups. Semi-supervised algorithms reflect a middle ground and are used when we have both labeled and unlabeled samples. Among the most widely used supervised learning algorithms for computational phenotyping are logistic regression, Bayesian networks, support vector machines (SVMs), decision trees, and random forests. More introductory and detailed description of supervised and unsupervised methods can be found in review papers such as Kotsiantis et al. [86] and Love et al. [87].

Regression methods have a long history of application for computational phenotyping [15, 28, 29]. Regression models adjust their parameters to maximize the conditional likelihood of the data. Further, regression models do not require a lot of effort in building or tuning, and the feature statistics derived from these regression models can be easily interpreted for meaningful insights.

In a study of identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis, Lin et al. [29] collected Concept Unique Identifiers (CUIs), Methotrexate (MTX) signatures, nearby words, and part-of-speech (POS) tags as features for an L2-regularized logistic regression. They obtained an F-measure of 83% in a performance evaluation. Liao et al. [88] implemented adaptive least absolute shrinkage and selection operator (LASSO) penalized logistic regression as classification algorithm to predict patients' probabilities of having Crohn's disease and achieved a PPV score of 98%. Both Lin's and Liao's methods experimented with a combination of features from structured EHR and NLP-processed features from clinical narratives. Their studies showed that the inclusion of NLP methods resulted in significantly improved performance for regression models. Due to the high dimensionality of features extracted from narratives, both methods applied regularized regressions.

Both Naive Bayes and Bayesian network classifiers are probabilistic classifiers [89] and work well with high-dimensional features. Unlike Bayesian networks, Naive Bayes doesn't require the inference of a dependency network and is more convenient in application when feature dimension is large. This is because Naive Bayes models assume that features are independent of one another whereas Bayesian networks allow for dependency among features. Besides their simplicity, Naive Bayes models are particularly useful for large datasets and are less prone to overfitting—sometimes outperforming highly sophisticated classification methods when sufficient data are available [90]. For example, Pakhomov et al. applied Naive Bayes to predict heart failure [91], using coded data (e.g., ICD-9, SNOMED) and a "bag of words" representation from clinical narratives as features. They chose Naive Bayes for their predictive algorithm due to its ability to process high-dimensional data. Their model achieved a sensitivity of 82% and a specificity of 98%. Similarly, Chase et al. [92] applied Naive Bayes for multiple sclerosis classifications and obtained an AUC score of 90%. Some studies have suggested that results obtained from logistic regression and Naive Bayes are comparable [93]. Copmared to logistic regression, the Naive Bayes classifier is capable of learning even in the presence of some missing values and relies less on missing data imputation [94, 95].

Table 1 Summarization and characterization of computational phenotyping systems. Abbreviations: CPT Current Procedural Terminology; CUI Concept Unique Identifier; cTAKES clinical Text Analysis and Knowledge Extraction System; ICD-9 International Classification of Diseases, ninth revision; NLP Natural Language Processing; UMLS Unified Medical Language System; TF-IDF Term Frequency-Inverse Document Frequency; HITEx Health Information Text Extraction; MedLEE Medical Language Extraction and Encoding System; KMCI KnowledgeMap Concept Indexer; NILE Narrative Information Linear Extraction.

Study	Assertion	Concept Extraction/ Concept Mapping	Data Source	Feature Generation
Aramaki et al. [96]	NA	Self-defined keywords	Narrative	Similarity score between sentences
Bejan, Vanderwende, et al. [97]	Section headers, self-defined features, NegEx, and ConTex	MetaMap	Restricted set of time order physician daily note	Uni-grams, bi-grams, UMLS concepts, assertion values associated with pneumonia expressions, statistical significance testing to rank features
Carroll et al. [27]	Modified form of NegEx in KMCI, section header	KMCI, MedEx for medication	Clinical notes, ICD-9	ICD-9, medication name, CUI, total note counts
Carroll et al. [35]	HITEx, Customized NegEx queries	HITEx	Diagnosis, billing, medication, procedural codes, physician text notes, discharge summaries, laboratory test results, radiology report	21 defined attributes from patients' narrative
Chapman et al. [98]	NA	SymText	X-ray reports	Pneumonia-related related concepts and its states from SymText
Chase et al. [92]	NA	MedLEE	Narrative	50 buckets representing pools of synonymous UMLS terms
Chen et al. [99]	NA	KMCI, SecTag, MedLEE, MedEx	Narrative, ICD-9, CPT	ICD-9, CPT, CUIs
Castro et al. [100]	Context dependent tokenizer in cTAKES	cTAKES	Radiology reports	Concepts, context dependent concepts, and concepts from cTAKES
Davis et al. [30]	Negation, word-sense disambiguation tool in KMCI	KMCI	ICD-9 codes, free text, and medications	ICD-9, CUIs, keywords
DeLisle et al. [101]	Customized rules, NegEx	Examined UMLS-supplied lexical variants/semantic types	Narrative, ICD-9, vital signs and orders for tests, imaging, and medications	186 UMLS associated with phenotype
DeLisle et al. [102]	NA	cTAKES	Chest imaging report ICD-9, encounter information, prescriptions	ICD-9, antibiotics medicine, hospital re-admission, binary variable of non-negative of chest imaging report
Fiszman et al. [75]	Self-defined rules	SymText	Chest x-ray reports	Set of augmented transition network grammars and a lexicon derived from the specialist Lexicon
Garla et al. [103]	cTAKES, YTEX, defined rules	Use cTAKES and YTEX to map concepts to UMLS and customized dictionary	Narrative and customized dictionary	Terms suggestive of benign/malignant lesions and UMLS concept in any liver-cancer related sentence
Gehrmann et al. [104]	cTAKES	cTAKES	Discharge summary	Concepts from cTAKES were transformed to continuous features using the TF-IDF

Haerian and Salmasian et al. [84]	Manual, and MedLEE	MedLEE map concepts to UMLS	Discharge summaries, ICD-9	MedLEE concepts were manually reviewed by a clinician. 31 codes were used
Herskovic et al. [105]	NA	MetaMap, SemRep	Narrative, biomedical literature	UMLS concept and UMLS relationship, semantic predication from biomedical literature
Lehman et al. [106]	NegEx	map to customized UMLS dictionary	Narrative	Manually selected UMLS concept
Li et al. [82]	NA	NA	Narrative, medication, lab results	Neonatologists manually reviewed 11 patients' notes and defined keywords and rules
Liao et al. [15]	Occurrence of concepts to indicate positive or negative of a sentence	HITEx	Provider notes, radiology reports, pathology reports, discharge summaries, operative reports, ICD-9, prescriptions	Concepts from HITEx, count of the concepts, binary variable to indicate occurrence of concepts.
Liao et al. [107]	NA	HITEx	ICD-9, CPT, lab results, narrative	Binary variable was created to indicate whether a concept was mentioned or not
Lin et al. [29]	cTAKES	cTAKES	Narrative, medication code, customized CUI	CUI, drug signatures (dosage, frequency), temporal features, nearby words, nearby POS tags
Luo et al. [108]	NA	Stanford Parser, Link Parser, ClearParser	Narrative	CUIs were used as nodes in the graph, syntactic dependencies among the concepts were used as edges in the graph
McCowan et al. [109]	NegEx	UMLS mapper	Pathology report	Map UMLS concepts to specific factors from the staging guidelines
Nguyen et al. [40]	NegEx, section heading	MEDTEX	Narrative	Concepts related to lung cancer resections (based on the AJCC 6th edition) were used
Ni et al. [54]	NegEx	cTAKES map to UMLS, SNOMED CT	Encounter data and clinical notes	Use concepts and encounter data. Predefine concepts from selection criteria, search for the hyponyms of query word
Nunes et al. [110]	Manual	NA	Narrative, ICD-9, lab results, demographics	Manual extract related terms and hyponyms and related words
Peissig et al. [111]	MedLEE	MedLEE	Narrative, ICD-9, CPT	UMLS concepts, ICD-9, CPT
Pineda et al. [112]	ConText	Topaz pipeline, map to UMLS	Narrative, lab test	Selected UMLS concepts and two lab test concepts
Posada et al. [113]	Section titles	MedLEE, keyword extraction, Question-Answer Feature Extraction	Psychiatric evaluation records	Count of keywords and concepts fall in nine defined categories as feature
Roque et al. [114]	NA	Simple sentence splitter split the text into smaller units	ICD-10, narrative	ICD-10, small units of sentences
Sauer et al. [85]	NA	Manual	Narrative	Experts reviewed notes and collected patterns to design

				extraction rules using regular expression
South et al. [115]	Negex	UMLS Metathesaurus	Narrative	NA
Teixeira et al. [116]	NegEx	MetaMap	Narrative, document count, medication, hypertension lab test related structured data	UMLS concepts, SNOMED-CT generated from narrative, ICD-9 code from structured data
Wang et al. [117]	NA	Standford parser	Clinical notes, comments, structured files	Constituent and dependency parsed from sentence
Ware et al. [80]	Self-Dev	NA	Narrative	Medication, treatment, word bigrams, numerical features, synonym list
Wei et al. [118]	cTAKES	cTAKES, map to SNOMED-CT	Narrative	SNOMED-CT concept, semantic type, node collapse concept
Wilke et al. [36]	NA	FreePharma	Narrative, ICD-9, laboratory data	NA
Xu et al. [34]	MedLEE	MedLEE, map to UMLS CUI	Narrative, ICD-9, CPT	UMLS concept, words of distance and direction (left vs. right)
Yu et al. [28]	NA	NILE, map to UMLS concept	ICD-9, Narrative	ICD-9, NLP features (counts of generic drug concept), number of notes for each patient
Zeng et al. [89]	HITEx (NexEx-2)	HITEx, map to UMLS	Narrative, ICD-9	N-word text fragments along with frequency, UMLS concept, smoking related sentences
Zhao et al. [119]	Self-defined	NA	PubMed knowledge, ICD-9, narrative	Selected concepts that are associated with pancreatic cancer

A Bayesian network consists of a directed acyclic graph whose node set contains random variables and whose edges represent relationships among the variables, and a conditional probability distribution of each node given each combination of values of its parents [120]. Bayesian networks have been used for reasoning in the presence of uncertainty and machine learning in many domains including biomedical informatics [121]. Chapman et al. [98] applied a Bayesian network inference model to predict pancreatic cancer using X-ray reports. In their experiments, a Bayesian network demonstrated high sensitivity 90% and specificity of 78%. Zhao et al. [119] applied a similar approach to identify pancreatic cancer. They developed a weighted Bayesian network with weights assigned to each node (feature). They also incorporated external knowledge from PubMed for scaling weights. Associations between each risk factor and pancreatic cancer were established using the output of NLP tools run on PubMed. Finally, they selected 20 risk factors as variables and fit them into a weighted Bayesian network model for pancreatic cancer prediction. Their results showed that this weighted Bayesian network achieved an AUC score of 91%, which had better performance than a traditional Bayesian network (81%). Compared to logistic regression or Naive Bayes methods, as a probabilistic formalism, Bayesian networks offer a better capacity to integrate heterogeneous knowledge in a single representation, which is particularly important in computational phenotyping because it complements the increasing availability of heterogeneous data sources [119]. A priori estimations can be taken into account in Bayesian network; this advantage allows one to incorporate known domain knowledge to increase model performances.

Clinical narratives are known to have high-dimensional feature spaces, few irrelevant features, and sparse instance vectors [122]. These problems were found to be well-addressed by SVMs [122]. In addition, SVMs have been recognized for their generalizability and are widely used for computational phenotyping [27, 89, 97, 103, 109, 123, 124]. In SVM models, a classifier is created by maximizing the margin between positive and negative examples [125]. Wei et al. [118] applied Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) to extract SNOMED CT concepts from clinical documents. The concepts were used to train a SVM for Type 2 Diabetes identification. Their algorithm achieved an F-measure of 95%. They concluded that concepts from the semantic type of disease or syndrome contain most important information for accurate phenotyping. Carroll et al. [27] implemented a SVM model for rheumatoid arthritis identification using a set of features from clinical narratives using the Knowledge Map Concept Identifier (KMCI) [126]. They demonstrated that a

SVM algorithm trained on these features outperformed a deterministic algorithm. Zeng et al. [89] trained a SVM model for principal diagnosis, co-morbidity, and smoking status identification. The features for the model were concepts extracted from discharge summaries and ICD-9 codes. The model achieved accuracies of 90% for smoking status, 87% for co-morbidity, and 82% for principal diagnoses. Chen et al. [99] applied active learning to a SVM classification algorithm to identify rheumatoid, colorectal cancer, and venous thromboembolism. Their results showed that active learning with a SVM could reduce annotated sample size while remaining relatively high performance. In the reviewed papers, SVMs constantly outperform other learning algorithms for computational phenotyping [27, 89, 99, 118, 127].

Kernel methods provide a structured way to extend the use of a linear algorithm to data that are not linearly separable by transforming the underlying feature space. The nonlinear transformation enables it to operate on high-dimensional data without explicitly computing the coordinates of the data in that space. SVMs are the most well-known learning algorithm using kernel based methods. Kotfila et al. [128] evaluated different SVM kernels' performances in identifying five diseases from unstructured medical notes. They found that SVMs with Gaussian radial basis function (RBF) kernels outperformed linear kernels. Zheng et al. [129] found that a SVM with RBF kernel exceeded non-kernel-based SVMs, decision trees, and perceptron for coreference resolution identification from the clinical narrative. In a study by Turner et al. [130], the authors tried to identify Systemic Lupus Erythematosus (SLE) from clinical notes. The authors concluded that a SVM with linear kernel outperformed radial basis function, polynomial, and sigmoid kernels. Good performance can be achieved in kernel methods with the appliance of statistical learning theory or Bayesian arguments. Linear methods are favored when there are many samples in a high dimensional input space. In contrast, for low-dimensional problems with many training instances, nonlinear kernel methods may be more favorable.Apart from the models mentioned above, researchers have explored other methods such as random forests [112], decision trees [100, 113, 131, 132], and the Longitudinal Gamma Poisson Shrinker [133, 134] for computational phenotyping. DeLisle et al. [102] implemented a conditional random field probabilistic classifier [135] to identify acute respiratory infections. They used structured data combined with narrative reports and demonstrated the inclusion of free text improved the PPV score by 20–70% while retaining sensitivities around 58-75%. Chapman et al. [98] applied decision trees, Bayesian networks, and an expert-crafted rule-based system to extract bacterial pneumonia from X-ray reports. The method using decision trees achieved an AUC score of 94%, and it is close to the other systems. Furthermore, semi-supervised methods have also been investigated for computational phenotyping [136, 137], which have the potential to significantly reduce the amount of labeling work and simultaneously retain high accuracy. Aramaki et al. [96] applied K-Nearest Neighbor classifier [138] based on the Okapi-BM25 similarity measure to extract patient smoking statuses from free text, and they achieved 89% accuracy in a performance evaluation. Carrero et al. [139] applied AdaBoost with Naive Bayes for text classification, and they achieved an F-measure of 72% using bigrams. Ni et al. [54] used TF-IDF similarity scores calculated from the feature vectors to identify a cohort of patients for clinical trial eligibility prescreening. Hybrid methods make use of more than one methods have also received increasing attention [138, 139], suggesting a promising direction for practical performance improvement.

For many data resources and domains, various models have been investigated, and some of them have achieved impressive success. However, a comprehensive understanding of the superior performance of a particular method over another for a specific domain remains an open challenge.

3.3 Unsupervised Learning

The time-consuming and labor-intensive process of obtaining labels for supervised learning algorithms limits their applicability to computational phenotyping. Another limitation of supervised learning is that it only looks for known characteristic patterns by designating a task and its outcome [86]. Unsupervised learning, on the other hand, can automatically classify phenotypes without extra annotations by experts [105, 140, 141]. Moreover, unsupervised learning searches for intrinsic patterns of data. Luo et al. [142] introduced subgraph augmented non-negative tensor factorization (SANTF) to cluster patients with lymphomas into three subtypes. After extracting atomic features (i.e., words) from narrative text, they implemented SANTF to mine relation features to cluster patients automatically. Their study demonstrated that NLP methods for unsupervised learning were able to achieve a decent accuracy (75%) and at the same time to discover latent subgroups. Roque et al. [114] extracted concepts from free text and mapped them to ICD-10 code. The ICD-10 code vector was used to represent each patient's profile and cosine similarity scores between each pair of ICD-10 vectors were obtained. Then, they applied hierarchical clustering to cluster those patients based on cosine similarity scores. As a result, they identified 26 clusters within 2,584 patients. They further analyzed the clinical characteristics of each cluster and concluded that NLP-based unsupervised learning was able to uncover the latent pattern of patient cohorts. Ho et al. [143] applied sparse non-negative tensor factorization on counts of normal and abnormal measurements obtained from EHR data for phenotype discovery. They identified multiple interpretable and concise phenotypes from a diverse EHR population, concluding that their methods were capable of characterizing and predicting a large number of diseases without supervision. Quan et al. [144] applied kernel-based pattern clustering and sentence parsing for interaction identification from narratives. In their application of protein-protein interaction, the unsupervised system achieved close performance to supervised methods.

Unsupervised learning has mitigated the laborious labeling work, thus making studies more scalable, and has the capability of finding novel phenotypes. However, interpretation of these new phenotypes requires domain expertise and remains challenging. Additionally, model performance in unsupervised learning is not yet as good as supervised learning. EHR-based unsupervised learning has frequently been applied on structured data [44, 45], but less frequently on narratives [142]. Further investigations on incorporating multiple data sources and at the same time maintaining or improving the performance are expected.

3.4 Deep Learning

Deep learning algorithms are good at finding intricate structures in high-dimensional data and have demonstrated good performance in natural language [145]. They have been adapted to learn vector representations of words for NLP-based phenotyping [112, 136], laying a foundation for computational phenotyping. Deep learning has been applied on various NLP applications, including semantic representation [146], semantic analysis [147, 148], information retrieval [149, 150], entity recognition [151, 152], relation extraction [153-156], and event detection [157, 158].

Beaulieu-Jones et al. [136] developed a neural network approach to construct phenotypes to classify patient disease status. The model obtained better performance than SVM, random forest, and decision tree models. They also claimed to successfully learn the structure of high-dimensional EHR data for phenotype stratification. Gehrmann et al. [104] compared convolutional neural networks (CNNs) to the traditional rule-based entity extraction systems using the cTAKES and logistic regression using n-gram features. They tested ten different phenotyping tasks using discharge summaries. The CNNs outperformed other phenotyping algorithms in the prediction of ten phenotypes, and they concluded that NLP-based deep learning methods improved the performance of patient phenotyping compared to other methods. Wu et al. [159] applied CNNs using a set of pre-trained embeddings on clinical text for named entity recognization. They found that their models outperformed the baseline of conditional random fields (CRF). Geraci et al. [160] applied deep neural networks to identify youth depression from unstructured text notes. The authors achieved a sensitivity of 93.5% and a specificity of 68%. Jagannatha et al. [161, 162] experimented with recurrent neural networks (RNNs), long short-term memory (LSTM), gated recurrent units (GRUs), bidirectional LSTMs, combinations of LSTMs with CRF, and CRF to extract clinical concepts from texts. They found that all variants of RNNs outperformed the CRF baseline. Lipton et al. [163] evaluated the performance of LSTM in phenotype prediction using multivariate time series clinical measurements. They concluded that their model outperformed logistic regression and multi-layer perceptron (MLP). They also concluded that the combination of LSTM and MLP had the best performance. Che et al. [164] also applied deep learning methods to study time series in ICU data. They introduced a prior-based Laplacian regularization process on the sigmoid layer that is based on medical ontologies and other structured knowledge. In addition, they developed an incremental training procedure to iteratively add neurons to the hidden layer. Then they applied causal inference techniques to analyze and interpret the hidden layer representations. They demonstrated that their proposed methods improved the performance of phenotype identification and that the model trains with faster convergence and better interpretation.

It is commonly known that unsupervised pre-training can improve deep learning performances and generalizability [165]. A generative deep learning algorithm that uses unsupervised methods can be applied to large unlabeled datasets, which has the potential to increases model generalizability [166]. Miotto et al. [167] applied a deep learning model called an auto-encoder as an unsupervised model to learn the latent representations for patients in order to predict their outcome and achieved better performance than principal component analysis. Due to the excellent model performance and good generalizability [168], using deep learning methods in conjunction with unsupervised methods is a promising approach in NLP-based computational phenotyping. Miotto et al. [169] introduced the framework of "deep patient". The method captures hierarchical regularities and dependencies in the data to create a vector for patient representation. This study showed that pre-processing data using a deep sequence of non-linear transformations can help better information embedding and information inference. Word2Vec [170] is an unsupervised artificial neural network (ANN) that has been developed to obtain vector representations of words when given large corpus and the representations are dependent on the context. For more details, we refer readers to a review [16] in recent advances on deep learning techniques for EHR analysis.

Even though deep learning methods present an opportunity to build phenotyping systems with good generalizability [171], a drawback of deep learning methods is their lack of interpretability. It can be difficult to understand how the features of the model arrive at predictions even though they can train a classifier with good performance [172].

4 MAKING NLP MORE EFFECTIVE

With numerous NLP methods available for computational phenotyping, it is practical to consider how to select more effective NLP methods or improve current NLP methods based on problem characteristics. This section reviews existing effort in these directions including model comparison, multi-modality data integration, entity recognition, and feature relation extractions.

4.1 Comparison of Models

Different computational phenotyping models vary in prediction accuracy and model generalizability. Comparison studies have been carried out to explore model performances. These comparison studies indicate algorithm performance differs based on specific conditions such as data sources, features, training data sizes, and target phenotypes.

In 1999, Wilcox et al. [173] conducted a study to investigate different algorithms' performances to extract clinical conditions from narratives. These algorithms were Naive Bayes, decision table, instance-based inducer, decision tree inducer MC4, decision tree inducer C5.0, and rule-discovery inducer CN2. Outputs of NLP algorithms were used as model features. They found MC4 and CN2 had the best performances while decision table performed the worst. Chapman et al. [98] tested rule-based method, Bayesian network, and decision tree for pneumonia detection using X-ray reports. The study showed that rule-based methods had slightly better performance (AUC score: 96%) than decision tree systems (AUC score: 94%) and Bayesian networks (AUC score: 95%).

Teixeira et al. [116] found random forests were superior to rule-based systems with a median AUC score of 98% when they were trying to identify hypertension using billing codes, medications, vitals, and concepts extracted from narratives. Pineda et al. [112] compared a Bayesian network classifier, Naive Bayes, a Bayesian network with the K2 algorithm, logistic regression, neural network, SVM, decision tree, and random forest for influenza detection. They concluded that all the machine learning classifiers had good performance with AUC score ranging from

88% to 93% and outperformed curated Bayesian network classifier, which had an AUC score of 80%.

Dumais et al. [132] compared the performances of SVM, Naive Bayes, Bayesian networks, decision trees, and rule-based systems in text classification. They concluded SVMs showed the best performance and noted that the training process is fast. Chen et al. [99] applied active learning to SVM classification, and their results showed that active learning with a SVM could reduce sample size needed. They concluded that semi-supervised learning, such as active learning, is efficient insofar as it reduces labeling cost.

Gehrmann et al. [104] compared convolutional neural networks (CNNs) to logistic regression and random forest model. They found CNNs had an improved performance compared to others and it can automatically learn the phrases associated with each patient phenotype, which reduced annotation complexity for clinical domain experts.

Among the compared methods, keyword search and rule-based systems often achieve good performance when such systems are well-designed and well-tuned. However, the construction of a keyword and rule list is laborious, making these systems difficult to scale. Supervised machine learning models have been favored for their capabilities of acquiring classification patterns and structures from data. The performance of supervised methods varies depending on the sample size, data resource type, number of data resources. Deep learning has also been favored for its better performance and generalizability. It has also been suggested that inclusion of more data resources can improve the model performances [174].

4.2 Combining Multiple Data Modalities

Computational phenotyping often involves multiple heterogeneous data sources in addition to structured data, such as clinical narratives, public databases, social media, biomedical literature [15, 88, 101, 111, 115, 175, 176]. Adding heterogeneous data has the benefit of providing complementary perspectives for computational phenotyping models [117]. Teixeira et al. [116] tested different combinations of ICD-9 codes, medications, vitals, and narrative documents as data resources for hypertension prediction. They found that model performance increases with the number of data resources regardless of the method used. They concluded that combination of multiple categories of information result in the best performances. The complete list of data sources utilized in the reviewed literature appears in

Table 1.

Liao et al. [15, 107] compared algorithms using ICD-9 codes alone to algorithms using a combination of structured data and NLP features. The results showed that the incorporation of NLP features improved algorithm performance significantly. Similarly, Nunes et al. [110] concluded that both structured data and clinical notes need to be considered to assess the occurrence of hypoglycemia among diabetes patients fully. Yu et al. [28] collected concepts from publicly available knowledge sources (e.g., Medscape, Wikipedia) and combined them with concepts extracted from narratives to predict rheumatoid arthritis (RA) and coronary artery (CAD) disease status. Their results showed that the combination of available public databases like Wikipedia and features derived from narratives could achieve high accuracy in RA and CAD prediction. Xu et al. [34] used ICD-9 codes, Current Procedural Terminology (CPT) codes, and colorectal cancer concepts to identify colorectal cancer. Zhao et al. [119] applied additional PubMed knowledge to weight the existing features.

The increasing trend of combining multiple data sources reflects the increased availability of EHR data and publicly available data [26]. Also, coupled with the increasing model complexities, there is a potential that more comprehensive data sources will be included for computational phenotyping. For example, one application developed by Gehrmann et al. [104] used CNNs to automatically learn the phrases associated with patients' phenotypes without task-specific rules or pre-defined keywords, which reduced the annotation effort for domain experts. As such, various data sources can be adopted for model training without too much human labor. However, regarding model generalizability, models and features based on narratives do not appear to be as portable as the ones based on structured EHR fields [116].

4.3 Entity Recognition and Relation Extraction

It is important to accurately recognize entities in clinical narratives as the extracted concepts are often used as features for models. Methods for feature learning vary from early-on manual selection to, more recently, machine learning methods. State-of-the-art named entity recognizers can automatically annotate text with high accuracy [177]. Bejan et al. [97, 123] implemented statistical feature selection, such as logistic regression with backward elimination to reduce feature dimensions. Wilcox et al. [173] tested machine learning algorithms with both expert-selected variables and automatically-selected variables by identifying top ranking predictive accuracy variables to classify six different diseases. Several studies, including those of Lehman et al., Luo et al., and Ghassemi et al. [106, 142, 178, 179], applied topic models and extended tensor-based topic models to learn better coherent features. Chen et al. [180] have applied an unsupervised system that is based on phrase chunking and distortional semantics to find features that are important to individual patients. Zhang et al. [181] have applied an unsupervised approach to extract named entities from biomedical text. Their model is a stepwise method, detecting entity boundaries and also classifying entities without pre-defined rules or annotated data. To do this, they assume that entities of same class tend to have similar vocabulary and context, which is called distributional semantics. Their model achieves a stable and competitive performance.

In addition to features, it is also critical to capture relations among features. Understanding these relations is important for knowledge representation and inference to augment structured knowledge bases [182, 183]. To date, a majority of the state-of-the-art methods for relation extraction are graph-based. Xu et al. [184] developed medication information extraction system (MedEx) to extract medications and relations between them. They applied the Kay Chart Parser [185] to parse sentences according to a self-defined grammar. In this way, they converted narratives to conceptual graph representations of medication relations. Using this graph representation, they were able to extract the association strength, frequencies, and routes. Representing medical concepts with graph nodes, Luo et al. [108]

augmented the Stanford Parser with UMLS-based concept recognition to generate graph representations for sentences in pathology reports. They then applied frequent subgraph mining to collect important semantic relations between medical concepts. The integration of named entity detection with relation extraction will produce end-to-end systems that can further automate the discovery and curation of novel biomedical knowledge. In addition, there is a trend towards increasingly unsupervised relation extraction, which is more adaptable across biomedical subdomains. Unsupervised methods have been investigated for feature relations too. Ciaramita et al. [186] presented an unsupervised model to learn semantic relations from text, hypothesizing that semantically related words co-occur more frequently. The model represented relations as syntactic dependency paths between ordered pairs of named entities. Relations were selected using the similarity scores associated with each class pair and dependency paths. Most recently, Alicante et al. [187] proposed using unsupervised methods for both entity and relation extraction from clinical notes. Clustering was applied to all the entity pairs for possible relations discovery.

5 FUTURE WORK

While notable progress has been made in computational phenotyping, challenges remain in developing generalizable, efficient, and effective models for accurate phenotype identification. Below we discuss these challenges and directions for future work.

5.1 Information heterogeneity in clinical narratives

Boland et al. [188] highlighted the heterogeneity apparent in clinical narratives due to the variance in physicians' expertise and behaviors. Different clinicians' perspectives can be quite different, and in practice they often are. Also, clinical narratives are often ungrammatical, incomplete with limited context, and contain a large number of abbreviations and acronyms [189], all of which make computational phenotyping challenging. Studies have applied UMLS or other external controlled vocabularies to recognize the various expressions of the same medical concept. However, performances of those external modules remain controversial [190, 191]. How to resolve the heterogeneity in clinical narratives remains an interesting topic.

5.2 Model generalizability

There is an ongoing trend of expanding generalizable algorithms to mine multiple diseases from different narratives. But these methods are still lacking in computational phenotyping [192, 193]. In addition, rule-based systems are one of the most prevalent methods for NLP-based computational phenotyping [3]. The intensive human labor required to adapt rules to a new system affects the model generalizability. Studies investigating algorithms that automatically mine rules are not yet available. Furthermore, even though statistical analysis and machine learning have provided alternative ways to automatically generate phenotypes, high dimensional feature spaces, data sparsity, and data imbalance remain impediments to the adoption of these methods [194]. Development of complete pipelines using various data sources for different phenotypes is one potential solution for generalizable computational phenotyping.

5.3 Model interpretability

More sophisticated models, such as convolutional neural networks, have the potential to automatically learn the phrases associated with each phenotype, which can reduce annotation complexity for clinical domain experts [104]. Using such models, one might be able to develop a system with good generalizability and have the availability to use multiple data sources. However, these same models tend to lack interpretability, which presents a problem that remains to be solved. Furthermore, meaningful interpretations of the novel phenotypes discovered in unsupervised clustering models remain one of the next big challenges in the field. Another promising direction is improving interpretation *while* retaining, or even improving, performance.

5.4 Characterizing the context of computational phenotyping

Clinical narratives contain patients' concerns, clinicians' assumptions, and patients' past medical histories. Clinicians also record diagnoses that are ruled out or symptoms that patients denied. Conditions, mentions, and feature relations can be extracted to better distinguish differential diagnoses. In computation phenotyping, generalized relation and event extraction, rather than binary relation classification, are expected to be a promising direction for future research; especially for the tasks of extracting clinical trial eligibility criteria [195], representing test results for automating diagnosis categorization [108], and building pharmacogenomic semantic networks [58], where the number of nodes is flexible, and the relation structure may not be entirely pre-specified due to the high complexity. To this end, graph methods are a promising class of algorithms and should be actively investigated [108, 142].

6 CONCLUSION

In this paper, we review the applications of NLP methods for EHR-based computational phenotyping, including the state-of-the-art NLP algorithms for this task. Our review shows that the keyword search, rule-based methods, and supervised machine learning-based NLP are the most widely used methods in the field. Well-designed keyword search and rule-based systems often show high accuracy. However, manually constructing keyword lists and rules results in problematically low generalizability and scalability for those methods.

Supervised classification has higher accuracy and is easy to train and test. However, the supervised classification methods require the training samples to be labeled, which can be labor intensive. To date, there is not a dominating method in the field; rather, model performances for the same type of methods may even vary depending on the data sources, data types, and sample sizes.

The combination of different data sources has the potential to improve model performance. Recently, unsupervised machine learning algorithms are gaining more attention because they require less human annotation and hold potential for finding novel phenotypes. Furthermore, new developments in machine learning methods, such as deep learning, have been increasingly adopted.

Finally, there is an emerging trend to extract relations between medical concepts as more expressive and powerful features. The extracted relations have been shown to increase algorithm performance significantly.

Despite these advances across multiple frontiers, there are many remaining challenges and opportunities for NLP-based computational phenotyping. These challenges include better model interpretability and generalizability, as well as proper characterization of feature relations in clinical narratives. These challenges will continuously shape the emerging landscape and provide research opportunities for NLP methods in EHR-based computational phenotyping.

ACKNOWLEDGMENT

This work was supported in part by NIH Grant 1R21LM012618-01, NLM Biomedical Informatics Training Grant 2T15 LM007092-22, and the Intel Science and Technology Center for Big Data. Corresponding author: Yuan Luo (yuan.luo@northwestern.edu)

REFERENCES

[1] R. L. Richesson *et al.*, "Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory," *Journal of the American Medical Informatics Association,* vol. 20, no. e2, pp. e226-e231, 2013.

[2] D. Blumenthal and M. Tavenner, "The "meaningful use" regulation for electronic health records," *N Engl J Med,* vol. 2010, no. 363, pp. 501-504, 2010.

[3] C. Shivade *et al.*, "A review of approaches to identifying patient phenotype cohorts using electronic health records," *Journal of the American Medical Informatics Association,* vol. 21, no. 2, pp. 221-230, 2013.

[4] W. H. Organization, "International Classification of Disease, 9th revision (ICD-9)," *Geneva: WHO Center for Classification of Disease,* 1977.

[5] W. H. Organization, "International statistical classification of diseases and health related problems, 10th revision," *Geneva: WHO,* 1992.

[6] C. Snomed, "Systematized nomenclature of medicine-clinical terms," *International Health Terminology Standards Development Organisation,* 2011.

[7] S. Liu, W. Ma, R. Moore, V. Ganesan, and S. Nelson, "RxNorm: prescription for electronic drug information exchange," *IT professional,* vol. 7, no. 5, pp. 17-23, 2005.

[8] A. W. Forrey *et al.*, "Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results," *Clinical Chemistry,* vol. 42, no. 1, pp. 81-90, 1996.

[9] P. Raghavan, J. L. Chen, E. Fosler-Lussier, and A. M. Lai, "How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?," *AMIA Summits on Translational Science Proceedings,* vol. 2014, p. 218, 2014.

[10] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage, "Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors," *Medical care,* vol. 43, no. 5, pp. 480-485, 2005.

[11] J. A. Singh, A. R. Holmgren, and S. Noorbaloochi, "Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis," *Arthritis Care & Research,* vol. 51, no. 6, pp. 952-957, 2004.

[12] K. J. O'malley, K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, and C. M. Ashton, "Measuring diagnoses: ICD code accuracy," *Health services research,* vol. 40, no. 5p2, pp. 1620-1639, 2005.

[13] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records," *Journal of the American Medical Informatics Association,* vol. 20, no. 1, pp. 117-121, 2012.

[14] T. Greenhalgh, "Narrative based medicine: narrative based medicine in an evidence based world," *BMJ: British Medical Journal,* vol. 318, no. 7179, p. 323, 1999.

[15] K. P. Liao *et al.*, "Electronic medical records for discovery research in rheumatoid arthritis," *Arthritis care & research,* vol. 62, no. 8, pp. 1120-1127, 2010.

[16] B. Shickel, P. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances on Deep Learning Techniques for Electronic Health Record (EHR) Analysis," *arXiv preprint arXiv:1706.03446,* 2017.

[17] J. McEntyre and D. Lipman, "PubMed: bridging the information gap," *Canadian Medical Association Journal,* vol. 164, no. 9, pp. 1317-1319, 2001.

[18] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," *PLoS biology,* vol. 2, no. 11, p. e309, 2004.

[19] K. Han, B. Park, H. Kim, J. Hong, and J. Park, "HPID: the human protein interaction database," *Bioinformatics,* vol. 20, no. 15, pp. 2466-2470, 2004.

[20] Y.-C. Fang, H.-C. Huang, and H.-F. Juan, "MeInfoText: associated gene methylation and cancer information from text mining," *BMC bioinformatics,* vol. 9, no. 1, p. 22, 2008.

[21] B. T. Alako *et al.*, "CoPub Mapper: mining MEDLINE based on search term co-publication," *BMC bioinformatics,* vol. 6, no. 1, p. 51, 2005.

[22] R. Hoffmann and A. Valencia, "Implementing the iHOP concept for navigation of biomedical literature," *Bioinformatics,* vol. 21, no. suppl_2, pp. ii252-ii258, 2005.

[23] F. Zhu *et al.*, "Biomedical text mining and its applications in cancer research," *Journal of biomedical informatics,* vol. 46, no. 2, pp. 200-211, 2013.

[24] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics,* vol. 6, no. 1, pp. 57-71, 2005.

[25] R. Kosala and H. Blockeel, "Web mining research: A survey," *ACM Sigkdd Explorations Newsletter,* vol. 2, no. 1, pp. 1-15, 2000.

[26] Y. Luo *et al.*, "Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review," *Drug Safety,* pp. 1-15, 2017.

[27] R. J. Carroll, A. E. Eyler, and J. C. Denny, "Naïve electronic health record phenotype identification for rheumatoid arthritis," in *AMIA annual symposium proceedings*, 2011, vol. 2011, p. 189: American Medical Informatics Association.

[28] S. Yu *et al.*, "Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources," *Journal of the American Medical Informatics Association,* vol. 22, no. 5, pp. 993-1000, 2015.

[29] C. Lin *et al.*, "Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record," *Journal of the American Medical Informatics Association,* vol. 22, no. e1, pp. e151-e161, 2014.

[30] M. F. Davis, S. Sriram, W. S. Bush, J. C. Denny, and J. L. Haines, "Automated extraction of clinical traits of multiple sclerosis in electronic medical records," *Journal of the American Medical Informatics Association,* vol. 20, no. e2, pp. e334-e340, 2013.

[31] N. L. Jain and C. Friedman, "Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports," in *Proceedings of the AMIA Annual Fall Symposium*, 1997, p. 829: American Medical Informatics Association.

[32] N. L. Jain, C. A. Knirsch, C. Friedman, and G. Hripcsak, "Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports," in *Proceedings of the AMIA Annual Fall Symposium*, 1996, p. 542: American Medical Informatics Association.

[33] G. Hripcsak, C. A. Knirsch, N. L. Jain, and A. Pablos-Mendez, "Automated tuberculosis detection," *Journal of the American Medical Informatics Association,* vol. 4, no. 5, pp. 376-381, 1997.

[34] H. Xu *et al.*, "Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases," in *AMIA Annual Symposium Proceedings*, 2011, vol. 2011, p. 1564: American Medical Informatics Association.

[35] R. J. Carroll *et al.*, "Portability of an algorithm to identify rheumatoid arthritis in electronic health records," *Journal of the American Medical Informatics Association,* vol. 19, no. e1, pp. e162-e169, 2012.

[36] R. A. Wilke *et al.*, "Use of an electronic medical record for the identification of research subjects with diabetes mellitus," *Clinical medicine & research,* vol. 5, no. 1, pp. 1-7, 2007.

[37] M. Panahiazar, V. Taslimitehrani, N. Pereira, and J. Pathak, "Using EHRs and machine learning for heart failure survival analysis," *Studies in health technology and informatics,* vol. 216, p. 40, 2015.

[38] Y. Wang *et al.*, "Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, 2015, pp. 2530-2533: IEEE.

[39] S. Lyalina, B. Percha, P. LePendu, S. V. Iyer, R. B. Altman, and N. H. Shah, "Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records," *Journal of the American Medical Informatics Association,* vol. 20, no. e2, pp. e297-e305, 2013.

[40] A. N. Nguyen *et al.*, "Symbolic rule-based classification of lung cancer stages from free-text pathology reports," *Journal of the American Medical Informatics Association,* vol. 17, no. 4, pp. 440-445, 2010.

[41] R. Haque *et al.*, "A hybrid approach to identify subsequent breast cancer using pathology and automated health information data," *Medical care,* vol. 53, no. 4, pp. 380-385, 2015.

[42] J. A. Strauss, C. R. Chao, M. L. Kwan, S. A. Ahmed, J. E. Schottinger, and V. P. Quinn, "Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm," *Journal of the American Medical Informatics Association,* vol. 20, no. 2, pp. 349-355, 2013.

[43] A. American Psychiatric, *Diagnostic and Statistical Manual of Mental Disorders.* American Psychiatric Publishing, 2013.

[44] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel, "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models," in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 2012, pp. 389-398: ACM.

[45] F. Doshi-Velez, Y. Ge, and I. Kohane, "Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis," *Pediatrics,* vol. 133, no. 1, pp. e54-e63, 2014.

[46] J. C. Ho *et al.*, "Limestone: High-throughput candidate phenotype generation via tensor factorization," *Journal of biomedical informatics,* vol. 52, pp. 199-211, 2014.

[47] Y. Luo, F. Wang, and P. Szolovits, "Tensor factorization toward precision medicine," *Briefings in Bioinformatics,* March 19, 2016 2016.

[48] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review,* vol. 51, no. 3, pp. 455-500, 2009.

[49] S. J. Shah *et al.*, "Phenomapping for novel classification of heart failure with preserved ejection fraction," *Circulation,* p. CIRCULATIONAHA. 114.010637, 2014.

[50] P. J. Embi, A. Jain, and C. M. Harris, "Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey," *BMC medical informatics and decision making,* vol. 8, no. 1, p. 13, 2008.

[51] A. J. Butte, D. A. Weinstein, and I. S. Kohane, "Enrolling patients into clinical trials faster using RealTime Recuiting," in *Proceedings of the AMIA Symposium*, 2000, p. 111: American Medical Informatics Association.

[52] P. J. Embi, A. Jain, J. Clark, and C. M. Harris, "Development of an electronic health record-based Clinical Trial Alert system to enhance recruitment at the point of care," in *AMIA Annual Symposium Proceedings*, 2005, vol. 2005, p. 231: American Medical Informatics Association.

[53] V. I. Petkov, L. T. Penberthy, B. A. Dahman, A. Poklepovic, C. W. Gillam, and J. H. McDermott, "Automated determination of metastases in unstructured radiology reports for eligibility screening in oncology clinical trials," *Experimental Biology and Medicine,* vol. 238, no. 12, pp. 1370-1378, 2013.

[54] Y. Ni *et al.*, "Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department," *Journal of the American Medical Informatics Association,* vol. 22, no. 1, pp. 166-178, 2014.

[55] S. R. Thadani, C. Weng, J. T. Bigger, J. F. Ennever, and D. Wajngurt, "Electronic screening improves efficiency in clinical trial recruitment," *Journal of the American Medical Informatics Association,* vol. 16, no. 6, pp. 869-873, 2009.

[56] D. L. Rubin, C. F. Thorn, T. E. Klein, and R. B. Altman, "A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge," *Journal of the American Medical Informatics Association,* vol. 12, no. 2, pp. 121-129, 2005.

[57] C. B. Ahlers, M. Fiszman, D. Demner-Fushman, F.-M. Lang, and T. C. Rindflesch, "Extracting semantic predications from Medline citations for pharmacogenomics," in *Pacific Symposium on Biocomputing*, 2007, vol. 12, pp. 209-220.

[58] A. Coulet, N. H. Shah, Y. Garten, M. Musen, and R. B. Altman, "Using text to build semantic networks for pharmacogenomics," *Journal of biomedical informatics,* vol. 43, no. 6, pp. 1009-1019, 2010.

[59] Y. Garten and R. B. Altman, "Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text," *BMC bioinformatics,* vol. 10, no. 2, p. S6, 2009.

[60] J. Delaney *et al.*, "Predicting clopidogrel response using DNA samples linked to an electronic health record," *Clinical Pharmacology & Therapeutics,* vol. 91, no. 2, pp. 257-263, 2012.

[61] B. Percha, Y. Garten, and R. B. Altman, "Discovery and explanation of drug-drug interactions via text mining," in *Pacific symposium on biocomputing. Pacific symposium on biocomputing*, 2012, p. 410: NIH Public Access.

[62] K. Haerian, D. Varn, S. Vaidya, L. Ena, H. Chase, and C. Friedman, "Detection of Pharmacovigilance-Related Adverse Events Using Electronic Health Records and Automated Methods," *Clinical Pharmacology & Therapeutics,* vol. 92, no. 2, pp. 228-234, 2012.

[63] E. Iqbal *et al.*, "Identification of adverse drug events from free text electronic patient records and information in a large mental health case register," *PloS one,* vol. 10, no. 8, p. e0134208, 2015.

[64] H. Zheng, H. Wang, H. Xu, Y. Wu, Z. Zhao, and F. Azuaje, "Linking biochemical pathways and networks to adverse drug reactions," *IEEE transactions on nanobioscience,* vol. 13, no. 2, pp. 131-137, 2014.

[65] W.-Q. Wei and J. C. Denny, "Extracting research-quality phenotypes from electronic health records to support precision medicine," *Genome medicine,* vol. 7, no. 1, p. 41, 2015.

[66] I. S. Kohane, "Using electronic health records to drive discovery in disease genomics," *Nature Reviews Genetics,* vol. 12, no. 6, pp. 417-428, 2011.

[67] K. P. Liao *et al.*, "Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non–rheumatoid arthritis controls," *Arthritis & Rheumatology,* vol. 65, no. 3, pp. 571-581, 2013.

[68] I. J. Kullo, J. Fan, J. Pathak, G. K. Savova, Z. Ali, and C. G. Chute, "Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease," *Journal of the American Medical Informatics Association,* vol. 17, no. 5, pp. 568-574, 2010.

[69] Y. Luo, F. S. Ahmad, and S. J. Shah, "Tensor factorization for precision medicine in heart failure with preserved ejection fraction," *Journal of Cardiovascular Translational Research,* pp. 1-8, 2017.

[70] W. S. Bush and J. H. Moore, "Genome-wide association studies," *PLoS computational biology,* vol. 8, no. 12, p. e1002822, 2012.

[71] J. C. Denny *et al.*, "PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations," *Bioinformatics,* vol. 26, no. 9, pp. 1205-1210, 2010.

[72] J. C. Denny *et al.*, "Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data," *Nature biotechnology,* vol. 31, no. 12, pp. 1102-1111, 2013.

[73] S. J. Hebbring, S. J. Schrodi, Z. Ye, Z. Zhou, D. Page, and M. H. Brilliant, "A PheWAS approach in studying HLA-DRB1* 1501," *Genes and immunity,* vol. 14, no. 3, pp. 187-191, 2013.

[74] M. D. Ritchie *et al.*, "Genome-and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk," *Circulation,* vol. 127, no. 13, pp. 1377-1385, 2013.

[75] M. Fiszman, W. W. Chapman, D. Aronsky, R. S. Evans, and P. J. Haug, "Automatic detection of acute bacterial pneumonia from chest X-ray reports," *Journal of the American Medical Informatics Association,* vol. 7, no. 6, pp. 593-604, 2000.

[76] S. M. Meystre and P. J. Haug, "Comparing natural language processing tools to extract medical problems from narrative text," in *AMIA annual symposium proceedings*, 2005, vol. 2005, p. 525: American Medical Informatics Association.

[77] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research,* vol. 32, no. suppl_1, pp. D267-D270, 2004.

[78] A. Wagholikar *et al.*, "Automated classification of limb fractures from free-text radiology reports using a clinician-informed gazetteer methodology," *The Australasian medical journal,* vol. 6, no. 5, p. 301, 2013.

[79] L. K. Wiley, J. D. Moretz, J. C. Denny, J. F. Peterson, and W. S. Bush, "Phenotyping adverse drug reactions: Statin-Related myotoxicity," *AMIA Summits on Translational Science Proceedings,* vol. 2015, p. 466, 2015.

[80] H. Ware, C. J. Mullett, and V. Jagannathan, "Natural language processing framework to assess clinical conditions," *Journal of the American Medical Informatics Association,* vol. 16, no. 4, pp. 585-589, 2009.

[81] C. Friedman, "Medlee-a medical language extraction and encoding system," *Columbia University, and Queens College of CUNY,* 1995.

[82] Q. Li *et al.*, "Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care," *Journal of the American Medical Informatics Association,* vol. 21, no. 5, pp. 776-784, 2014.

[83] T. N. Raju, G. Suresh, and R. D. Higgins, "Patient safety in the context of neonatal intensive care: research and educational opportunities," *Pediatric research,* vol. 70, no. 1, p. 109, 2011.

[84] K. Haerian, H. Salmasian, and C. Friedman, "Methods for identifying suicide or suicidal ideation in EHRs," in *AMIA Annual Symposium Proceedings*, 2012, vol. 2012, p. 1244: American Medical Informatics Association.

[85] B. C. Sauer *et al.*, "Performance of a Natural Language Processing (NLP) tool to extract pulmonary function test (PFT) reports from structured and semistructured Veteran Affairs (VA) data," *eGEMs,* vol. 4, no. 1, 2016.

[86] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," ed, 2007.

[87] B. C. Love, "Comparing supervised and unsupervised category learning," *Psychonomic bulletin & review,* vol. 9, no. 4, pp. 829-835, 2002.

[88] K. P. Liao *et al.*, "Development of phenotype algorithms using electronic medical records and incorporating natural language processing," *bmj,* vol. 350, p. h1885, 2015.

[89] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system," *BMC medical informatics and decision making,* vol. 6, no. 1, p. 30, 2006.

[90] J. Huang, J. Lu, and C. X. Ling, "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 2003, pp. 553-556: IEEE.

[91] S. Pakhomov, S. A. Weston, S. J. Jacobsen, C. G. Chute, R. Meverden, and V. L. Roger, "Electronic medical records for clinical research: application to the identification of heart failure," *Am J Manag Care,* vol. 13, no. 6 Part 1, pp. 281-288, 2007.

[92] H. S. Chase, L. R. Mitrani, G. G. Lu, and D. J. Fulgieri, "Early recognition of multiple sclerosis using natural language processing of the electronic health record," *BMC medical informatics and decision making,* vol. 17, no. 1, p. 24, 2017.

[93] P. Sebastiani, N. Solovieff, and J. X. Sun, "Naïve Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: not so different after all!," *Frontiers in genetics,* vol. 3, 2012.

[94] Y. Luo, P. Szolovits, A. S. Dighe, and J. M. Baron, "Using Machine Learning to Predict Laboratory Test Results," *American Journal of Clinical Pathology,* vol. 145, no. 6, pp. 778-788, 2016.

[95] Y. Luo, P. Szolovits, A. S. Dighe, and J. M. Baron, "3D-MICE: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data," (in eng), *J Am Med Inform Assoc,* Nov 30 2017.

[96] E. Aramaki, T. Imai, K. Miyo, and K. Ohe, "Patient status classification by using rule based sentence extraction and BM25 kNN-based classifier," in *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 2006.

[97] C. A. Bejan, F. Xia, L. Vanderwende, M. M. Wurfel, and M. Yetisgen-Yildiz, "Pneumonia identification using statistical feature selection," *Journal of the American Medical Informatics Association,* vol. 19, no. 5, pp. 817-823, 2012.

[98] W. W. Chapman, M. Fizman, B. E. Chapman, and P. J. Haug, "A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia," *Journal of biomedical informatics,* vol. 34, no. 1, pp. 4-14, 2001.

[99] Y. Chen *et al.*, "Applying active learning to high-throughput phenotyping algorithms for electronic health records data," *Journal of the American Medical Informatics Association,* vol. 20, no. e2, pp. e253-e259, 2013.

[100] S. M. Castro *et al.*, "Automated annotation and classification of BI-RADS assessment from radiology reports," *Journal of Biomedical Informatics,* vol. 69, pp. 177-187, 2017.

[101] S. DeLisle *et al.*, "Combining free text and structured electronic medical record entries to detect acute respiratory infections," *PloS one,* vol. 5, no. 10, p. e13377, 2010.

[102] S. DeLisle *et al.*, "Using the electronic medical record to identify community-acquired pneumonia: toward a replicable automated strategy," *PLoS One,* vol. 8, no. 8, p. e70944, 2013.

[103] V. Garla, C. Taylor, and C. Brandt, "Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management," *Journal of biomedical informatics,* vol. 46, no. 5, pp. 869-875, 2013.

[104] S. Gehrmann *et al.*, "Comparing Rule-Based and Deep Learning Models for Patient Phenotyping," *arXiv preprint arXiv:1703.08705,* 2017.

[105] J. R. Herskovic, D. Subramanian, T. Cohen, P. A. Bozzo-Silva, C. F. Bearden, and E. V. Bernstam, "Graph-based signal integration for high-throughput phenotyping," *BMC bioinformatics,* vol. 13, no. 13, p. S2, 2012.

[106] L.-w. Lehman, M. Saeed, W. Long, J. Lee, and R. Mark, "Risk stratification of ICU patients using topic models inferred from unstructured progress notes," in *AMIA annual symposium proceedings*, 2012, vol. 2012, p. 505: American Medical Informatics Association.

[107] K. P. Liao *et al.*, "Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts," *PloS one,* vol. 10, no. 8, p. e0136651, 2015.

[108] Y. Luo, A. R. Sohani, E. P. Hochberg, and P. Szolovits, "Automatic lymphoma classification with sentence subgraph mining from pathology reports," *Journal of the American Medical Informatics Association,* vol. 21, no. 5, pp. 824-832, 2014.

[109] I. A. McCowan *et al.*, "Collection of cancer stage data by classifying free-text medical reports," *Journal of the American Medical Informatics Association,* vol. 14, no. 6, pp. 736-745, 2007.

[110] A. P. Nunes *et al.*, "Assessing occurrence of hypoglycemia and its severity from electronic health records of patients with type 2 diabetes mellitus," *Diabetes research and clinical practice,* vol. 121, pp. 192-203, 2016.

[111] P. L. Peissig *et al.*, "Importance of multi-modal approaches to effectively identify cataract cases from electronic health records," *Journal of the American Medical Informatics Association,* vol. 19, no. 2, pp. 225-234, 2012.

[112] A. L. Pineda, Y. Ye, S. Visweswaran, G. F. Cooper, M. M. Wagner, and F. R. Tsui, "Comparison of machine learning classifiers for influenza detection from emergency department free-text reports," *Journal of biomedical informatics,* vol. 58, pp. 60-69, 2015.

[113] J. D. Posada *et al.*, "Predictive Modeling for Classification of Positive Valence System Symptom Severity from Initial Psychiatric Evaluation Records," *Journal of Biomedical Informatics,* 2017.

[114] F. S. Roque *et al.*, "Using electronic patient records to discover disease correlations and stratify patient cohorts," *PLoS computational biology,* vol. 7, no. 8, p. e1002141, 2011.

[115] B. R. South, S. Shen, W. W. Chapman, S. Delisle, M. H. Samore, and A. V. Gundlapalli, "Analysis of false positive errors of an acute respiratory infection text classifier due to contextual features," *Summit on Translational Bioinformatics,* vol. 2010, p. 56, 2010.

[116] P. L. Teixeira *et al.*, "Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals," *Journal of the American Medical Informatics Association,* vol. 24, no. 1, pp. 162-171, 2016.

[117] Y. Wang, E. S. Chen, S. Pakhomov, E. Lindemann, and G. B. Melton, "Investigating Longitudinal Tobacco Use Information from Social History and Clinical Notes in the Electronic Health Record," in *AMIA Annual Symposium Proceedings*, 2016, vol. 2016, p. 1209: American Medical Informatics Association.

[118] W.-Q. Wei, C. Tao, G. Jiang, and C. G. Chute, "A high throughput semantic concept frequency based approach for patient identification: a case study using type 2 diabetes mellitus clinical notes," in *AMIA annual symposium proceedings*, 2010, vol. 2010, p. 857: American Medical Informatics Association.

[119] D. Zhao and C. Weng, "Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction," *Journal of biomedical informatics,* vol. 44, no. 5, pp. 859-868, 2011.

[120] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning,* vol. 29, no. 2-3, pp. 131-163, 1997.

[121] Z. Zeng, X. Jiang, and R. Neapolitan, "Discovering causal interactions using Bayesian network scoring and information gain," *BMC bioinformatics,* vol. 17, no. 1, p. 1, 2016.

[122] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine learning: ECML-98,* pp. 137-142, 1998.

[123] C. A. Bejan, L. Vanderwende, M. M. Wurfel, and M. Yetisgen-Yildiz, "Assessing pneumonia identification from time-ordered narrative reports," in *AMIA Annual Symposium Proceedings*, 2012, vol. 2012, p. 1119: American Medical Informatics Association.

[124] Z. Zeng *et al.*, "Contralateral Breast Cancer Event Detection Using Nature Language Processing."

[125] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *The annals of statistics,* vol. 26, no. 5, pp. 1651-1686, 1998.

[126] J. C. Denny, J. D. Smithers, R. A. Miller, and A. Spickard III, "“Understanding” medical school curriculum content using KnowledgeMap," *Journal of the American Medical Informatics Association,* vol. 10, no. 4, pp. 351-362, 2003.

[127] Y. Goldberg and M. Elhadad, "splitSVM: fast, space-efficient, non-heuristic, polynomial kernel computation for NLP applications," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, 2008, pp. 237-240: Association for Computational Linguistics.

[128] C. Kotfila and Ö. Uzuner, "A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases," *Journal of biomedical informatics,* vol. 58, pp. S92-S102, 2015.

[129] J. Zheng, W. W. Chapman, T. A. Miller, C. Lin, R. S. Crowley, and G. K. Savova, "A system for coreference resolution for the clinical narrative," *Journal of the American Medical Informatics Association,* vol. 19, no. 4, pp. 660-667, 2012.

[130] C. A. Turner *et al.*, "Word2Vec inversion and traditional text classifiers for phenotyping lupus," *BMC medical informatics and decision making,* vol. 17, no. 1, p. 126, 2017.

[131] P. Mukherjee *et al.*, "NegAIT: A new parser for medical text simplification using morphological, sentential and double negation," *Journal of Biomedical Informatics,* vol. 69, pp. 55-62, 2017.

[132] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the seventh international conference on Information and knowledge management*, 1998, pp. 148-155: ACM.

[133] C. Ferrajolo *et al.*, "Idiopathic acute liver injury in paediatric outpatients: incidence and signal detection in two European countries," *Drug safety,* vol. 36, no. 10, p. 1007, 2013.

[134] C. Ferrajolo *et al.*, "Signal detection of potentially drug-induced acute liver injury in children using a multi-country healthcare database network," *Drug safety,* vol. 37, no. 2, p. 99, 2014.

[135] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[136] B. K. Beaulieu-Jones and C. S. Greene, "Semi-supervised learning of the electronic health record for phenotype stratification," *Journal of biomedical informatics,* vol. 64, pp. 168-178, 2016.

[137] Z. Wang, A. D. Shah, A. R. Tate, S. Denaxas, J. Shawe-Taylor, and H. Hemingway, "Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning," *PLoS One,* vol. 7, no. 1, p. e30412, 2012.

[138] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory,* vol. 13, no. 1, pp. 21-27, 1967.

[139] F. Carrero, J. G. Hidalgo, E. Puertas, M. Maña, and J. Mata, "Quick prototyping of high performance text classifiers," in *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 2006.

[140] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data," *PloS one,* vol. 8, no. 6, p. e66341, 2013.

[141] W.-J. Guan, M. Jiang, Y.-H. Gao, R.-C. Chen, and N.-S. Zhong, "In Reply: Towards precision medicine: phenotyping bronchiectasis with unsupervised learning technique," *The International Journal of Tuberculosis and Lung Disease,* vol. 20, no. 5, pp. 710-710, 2016.

[142] Y. Luo, Y. Xin, E. Hochberg, R. Joshi, O. Uzuner, and P. Szolovits, "Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text," *Journal of the American Medical Informatics Association,* p. ocv016, 2015.

[143] J. C. Ho, J. Ghosh, and J. Sun, "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 115-124: ACM.

[144] C. Quan, M. Wang, and F. Ren, "An unsupervised text mining method for relation extraction from biomedical literature," *PloS one,* vol. 9, no. 7, p. e102039, 2014.

[145] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research,* vol. 12, no. Aug, pp. 2493-2537, 2011.

[146] S. W.-t. Yih, X. He, and C. Meek, "Semantic parsing for single-relation question answering," 2014.

[147] J. Zhou and W. Xu, "End-to-end learning of semantic role labeling using recurrent neural networks," in *ACL (1)*, 2015, pp. 1127-1137.

[148] A. Mazalov, B. Martins, and D. Matos, "Spatial role labeling with convolutional neural networks," in *Proceedings of the 9th Workshop on Geographic Information Retrieval*, 2015, p. 12: ACM.

[149] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 373-374: ACM.

[150] A. Severyn and A. Moschitti, "Learning to rank short text pairs with convolutional deep neural networks," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 373-382: ACM.

[151] H. Huang, L. Heck, and H. Ji, "Leveraging deep neural networks and knowledge graphs for entity disambiguation," *arXiv preprint arXiv:1504.07678,* 2015.

[152] T. H. Nguyen, A. Sil, G. Dinu, and R. Florian, "Toward Mention Detection Robustness with Recurrent Neural Networks," *arXiv preprint arXiv:1602.07749,* 2016.

[153] T. H. Nguyen and R. Grishman, "Combining neural networks and log-linear models to improve relation extraction," *arXiv preprint arXiv:1511.05926,* 2015.

[154] X. Yan, L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin, "Classifying relations via long short term memory networks along shortest dependency path," *arXiv preprint arXiv:1508.03720,* 2015.

[155] Y. Luo, "Recurrent Neural Networks for Classifying Relations in Clinical Notes," *Journal of Biomedical Informatics,* 2017.

[156] Y. Luo, Y. Cheng, O. Uzuner, P. Szolovits, and J. Starren, "Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes," *Journal of the American Medical Informatics Association,* no. ocx090, 2017.

[157] P. Dasigi and E. H. Hovy, "Modeling Newswire Events using Neural Networks for Anomaly Detection," in *COLING*, 2014, pp. 1414-1422.

[158] T. H. Nguyen and R. Grishman, "Event Detection and Domain Adaptation with Convolutional Neural Networks," in *ACL (2)*, 2015, pp. 365-371.

[159] Y. Wu, M. Jiang, J. Lei, and H. Xu, "Named entity recognition in Chinese clinical text using deep neural network," *Studies in health technology and informatics,* vol. 216, p. 624, 2015.

[160] J. Geraci, P. Wilansky, V. de Luca, A. Roy, J. L. Kennedy, and J. Strauss, "Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression," *Evidence-based mental health,* vol. 20, no. 3, pp. 83-87, 2017.

[161] A. N. Jagannatha and H. Yu, "Structured prediction models for RNN based sequence labeling in clinical text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing,* 2016, vol. 2016, p. 856: NIH Public Access.

[162] A. N. Jagannatha and H. Yu, "Bidirectional RNN for medical event detection in electronic health records," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting,* 2016, vol. 2016, p. 473: NIH Public Access.

[163] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with LSTM recurrent neural networks," *arXiv preprint arXiv:1511.03677,* 2015.

[164] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, "Deep computational phenotyping," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2015, pp. 507-516: ACM.

[165] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research,* vol. 11, no. Feb, pp. 625-660, 2010.

[166] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation,* vol. 18, no. 7, pp. 1527-1554, 2006.

[167] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports,* vol. 6, 2016.

[168] M. Yousefi-Azar and L. Hamey, "Text summarization using unsupervised deep learning," *Expert Systems with Applications,* vol. 68, pp. 93-105, 2017.

[169] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports,* vol. 6, p. 26094, 2016.

[170] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781,* 2013.

[171] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882,* 2014.

[172] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," (in eng), *Brief Bioinform,* May 06 2017.

[173] A. Wilcox and G. Hripcsak, "Classification algorithms applied to narrative reports," in *Proceedings of the AMIA Symposium*, 1999, p. 455: American Medical Informatics Association.

[174] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems,* vol. 24, no. 2, pp. 8-12, 2009.

[175] Y. Luo, G. Riedlinger, and P. Szolovits, "Text mining in cancer gene and pathway prioritization," *Cancer informatics,* no. Suppl. 1, p. 69, 2014.

[176] Y. Luo, Y. Xin, R. Joshi, L. Celi, and P. Szolovits, "Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016.

[177] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association,* vol. 18, no. 5, pp. 552-556, 2011.

[178] M. Ghassemi *et al.*, "Unfolding physiological state: mortality modelling in intensive care units," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 75-84: ACM.

[179] M. Ghassemi *et al.*, "A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data," in *AAAI*, 2015, pp. 446-453.

[180] J. Chen and H. Yu, "Unsupervised ensemble ranking of terms in electronic health record notes based on their importance to patients," *Journal of Biomedical Informatics,* vol. 68, pp. 121-131, 2017.

[181] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *Journal of biomedical informatics,* vol. 46, no. 6, pp. 1088-1098, 2013.

[182] Y. Luo, Ö. Uzuner, and P. Szolovits, "Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations," *Briefings in bioinformatics,* vol. 18, no. 1, pp. 160-178, 2016.

[183] Y. Luo and O. Uzuner, "Semi-Supervised Learning to Identify UMLS Semantic Relations," in *AMIA Joint Summits on Translational Science*, 2014: American Medical Informatics Association.

[184] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "MedEx: a medication information extraction system for clinical narratives," *Journal of the American Medical Informatics Association,* vol. 17, no. 1, pp. 19-24, 2010.

[185] M. Kay, "Algorithm schemata and data structures in syntactic processing," *Technical Report CSL80-12,* 1980.

[186] M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric, and I. Rojas, "Unsupervised learning of semantic relations between concepts of a molecular biology ontology," in *IJCAI*, 2005, pp. 659-664.

[187] A. Alicante, A. Corazza, F. Isgrò, and S. Silvestri, "Unsupervised entity and relation extraction from clinical records in Italian," *Computers in biology and medicine,* vol. 72, pp. 263-275, 2016.

[188] M. R. Boland, G. Hripcsak, Y. Shen, W. K. Chung, and C. Weng, "Defining a comprehensive verotype using electronic health records for personalized medicine," *Journal of the American Medical Informatics Association,* pp. e232-e238, 2013.

[189] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association,* vol. 18, no. 5, pp. 544-551, 2011.

[190] W. Hersh, S. Price, and L. Donohoe, "Assessing thesaurus-based query expansion using the UMLS Metathesaurus," in *Proceedings of the AMIA Symposium*, 2000, p. 344: American Medical Informatics Association.

[191] A. Passos and J. Wainer, "Wordnet-based metrics do not seem to help document clustering," in *International Workshop on Web and Text Intelligence (WTI-2009)*, 2009.

[192] L. Cui, S. D. Lhatoo, G.-Q. Zhang, S. S. Sahoo, and A. Bozorgi, "EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification," in *AMIA*, 2012.

[193] D. B. Aronow, F. Fangfang, and W. B. Croft, "Ad hoc classification of radiology reports," *Journal of the American Medical Informatics Association,* vol. 6, no. 5, pp. 393-411, 1999.

[194] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[195] C. Weng, X. Wu, Z. Luo, M. R. Boland, D. Theodoratos, and S. B. Johnson, "EliXR: an approach to eligibility criteria extraction and representation," *Journal of the American Medical Informatics Association,* vol. 18, no. Supplement_1, pp. i116-i124, 2011.

Zexian Zeng received both his master degrees in Industrial and Systems Engineering and Computer Science from University of Wisconsin-Madison in 2014. Currently, he is working towards the Ph.D. degree at the Department of Preventive Medicine, Northwestern University Feinberg School of Medicine. His research interests are in natural language processing and cancer genomics.

Yu Deng earned her bachelor degree in Biotechnology from Northeast Normal University in 2014, China. Currently, she is working towards her PhD degree in Biomedical Informatics, Northwestern University. She works on the development of mathematics and computer methods for dynamic risk prediction. She is also interested in disease sub-phenotype discovery using clustering methods. She is a member of American Medical Informatics Association since 2016.

Xiaoyu Li received a BA in Sociology from Tsinghua University in 2011, an MS in sociology from University of Wisconsin-Madison in 2013, and an MS in Biostatistics and a ScD in Social and Behavioral Sciences from Harvard T.H. Chan School of Public Health in 2017. She is currently a postdoctoral research fellow at Harvard T.H. Chan School of Public Health and Brigham and Women's Hospital. She is interested in data science methods and contextual determinants of population health. She is a reviewer for the journal of Social Science and Medicine and a member of the Society for Epidemiologic Research.

Tristan Naumann is a Ph.D. candidate in Electrical Engineering and Computer Science at MIT working with Professor Peter Szolovits in CSAIL’s Clinical Decision-Making group. His research includes exploring relationships in complex, unstructured healthcare data using natural language processing and unsupervised learning techniques. He has been an organizer for workshops and datathon events, which bring together participants with diverse backgrounds in order to address biomedical and clinical questions in a manner that is reliable and reproducible.

Yuan Luo is an assistant professor in the Department of Preventive Medicine at Northwestern University Feinberg School of Medicine. He received his PhD in Computer Science from Massachusetts Institute of Technology. He served on the student editorial board of Journal of American Medical Informatics Association. His research interests include machine learning, natural language processing, time series analysis, computational genomics, with a focus on biomedical applications. Dr. Luo is the recipient of the inaugural Doctoral Dissertation Award Honorable Mention by American Medical Informatics Association (AMIA) in 2017.