Machine learning approaches for electronic health records phenotyping: A methodical review

Siyue Yang[1], Paul Varghese[2], Ellen Stephenson[3], Karen Tu[3], and Jessica Gronsbell[1,3*]

[1]*Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada*

[2]*Verily Life Sciences, Cambridge, MA, USA*

[3]*Department of Family & Community Medicine, University of Toronto, Toronto, ON, Canada*

Correspondence to:

Jessica Gronsbell

Postal address: Department of Statistical Sciences, University of Toronto, 700 University Ave.

Toronto, ON M5G 1Z5, Canada

Email: j.gronsbell@utoronto.ca

Keywords: Electronic Health Records; Phenotyping; Cohort Identification; Machine Learning

Objective Accurate and rapid methods for phenotyping are a prerequisite to realizing the potential of electronic health records (EHRs) data for clinical and translational research. This study reviews the literature on machine learning (ML) approaches for phenotyping with respect to the phenotypes considered, the data sources and methods used, and the contributions within the wider context of EHR-based research.

Materials and Methods We searched for relevant articles in PubMed and Web of Science published between January 1, 2018 and April 14, 2022. After screening, we collected data on 52 variables across 106 selected articles.

Results ML-based methods were developed for 156 unique phenotypes, primarily using EHR data from a single institution or health system. 72 of 106 articles leveraged unstructured data in clinical notes. In terms of methodology, supervised learning is the most prevalent ML paradigm (n = 64, 60.4%), with half of the articles employing deep learning. Semi-supervised and weakly-supervised approaches were applied to reduce the burden of obtaining gold-standard labeled data (n = 21, 19.8%), while unsupervised learning was used for phenotype discovery (n = 20, 18.9%). Federated learning has been applied to develop algorithms across multiple institutions while preserving data privacy (n = 2, 1.9%).

Discussion While the use of ML for phenotyping is growing, most articles applied traditional supervised ML to characterize the presence of common, chronic conditions.

Conclusion Continued research in ML-based methods is warranted, with particular attention to the development of advanced methods for complex phenotypes and standards for reporting and evaluating phenotyping algorithms.

INTRODUCTION

Electronic health records (EHRs) are becoming a central data source for clinical and translational research due to their availability, scale, and granularity [1]. In recent years, EHR data has been used to support discovery in disease genomics, to facilitate rapid and more inclusive clinical trials recruitment, and to derive real-world predictive models for understudied and emerging diseases [2–5]. EHR data is also positioned to be a key data source for the development of personalized disease prevention and treatment strategies [6,7]. This myriad of opportunities rests on the ability to accurately and rapidly characterize a wide variety of patient phenotypes, such as the presence of a disease, treatment response, or lifestyle factor, in order to identify patient populations of interest and to define relevant outcome, exposure, and clinical variables. However, precise phenotype information is not readily available in patient records and presents a major barrier to realizing the potential of EHR-driven research [8–10].

Historically, phenotypes have been inferred from rule-based algorithms using expert defined criteria from care guidelines and/or knowledge of clinical practice [11]. Because documentation varies substantially across phenotypes, providers, and institutions, developing a sufficient set of rules that encode how a condition may be documented in patient records is prohibitively resource intensive [12,13]. Moreover, the scope of rule-based algorithms is severely limited, particularly for more complex phenotypes recorded primarily in clinical notes [14–16]. In recent years, there has been significant interest in phenotyping algorithms derived from machine learning (ML) models to improve the efficiency and accuracy of algorithm development across a broad range of phenotypes and healthcare settings [14,17]. In contrast to rule-based methods, ML approaches

aggregate multiple sources of information available in patient records to improve phenotype characterization in a more automated and generalizable fashion [18–20].

To evaluate recent advances in ML-based phenotyping, this work summarizes the literature with respect to the phenotypes considered, the data sources and methods used, and the primary contributions within the broader landscape of EHR and ML research (**Figure 1**). Through a methodical review of the literature and analysis of 52 variables recorded across 106 articles, our key findings include:

- Most phenotyping algorithms have been developed for common, chronic conditions such as diabetes and cancer using data from a single institution or healthcare system.
- A majority of articles utilize traditional supervised learning methods, but there is increasing use of deep learning methods and approaches that leverage clinical notes.
- Semi-supervised and weakly-supervised methods have reduced the burden of obtaining gold-standard labeled data while maintaining accuracy comparable to supervised learning.
- Federated learning is beginning to enable multi-institution algorithm development while respecting data privacy constraints.
- There are inconsistencies in the reporting of study settings and code is rarely released with published papers.

MATERIALS AND METHODS

We closely followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines when conducting our review of the literature. **Figure 2** illustrates the search strategy.

Search strategy

We searched for full-text articles in PubMed and Web of Science published between January 1, 2018 and April 14, 2022. The query was initially conducted on December 23, 2021 and updated on April 14, 2022. The range of publication year was specified to not overlap with existing reviews [10,11,14,17,19–22]. Existing reviews are compared with this article in Supplementary Material **Table S1**. We limited our search to (1) *Journal of American Medical Informatics Association* (JAMIA), (2) *JAMIA Open*, (3) *Journal of Biomedical Informatics* (JBI), (4) *PloS One,* and (5) *Proceedings of the American Medical Informatics Association's Annual Symposium* (AMIA) to target articles focused primarily on phenotyping, rather than clinical applications using EHR data. This strategy was originally proposed in [11] using a manual review of all articles published in (1, 3-5) and later extended to a PubMed query in [14]. We increased the scope of our search by including Web of Science and added additional search strings related to ML [23]. Due to the broad definition of phenotyping, our search was designed to capture articles published in English that (i) utilized ML or considered phenotyping and (ii) used EHRs as the primary data source. The queries for each database are provided in Supplementary Material **Table S2**.

Study selection

Title and abstract screening

After removing duplicates, articles were retrieved and underwent title and abstract screening by two authors (S.Y. and J.G.). A third author (P.V.) resolved disagreements. Articles were excluded if they (i) were reviews, perspectives, or editorials, (ii) did not use EHRs as a primary data source, (iii) did not use ML methods, or (iv) did not consider phenotyping. Consistent with recent literature, we used an inclusive definition of phenotyping as any procedure that makes a characterization about a patient (eg. disease status, treatment response, lifestyle factor) based on the information in their health record [24]. A detailed list of article exclusions is provided in Supplementary Material **Table S3**.

Full-text review

One author (S.Y.) reviewed the full-text of articles. Another author (J.G.) verified the information from a full-text review. After excluding papers that did not focus on ML approaches for EHR-based phenotyping, 106 papers were selected for this review. A complete list of selected articles is in Supplementary Material **Table S4**.

Data collection

During full-text review, we extracted information related to (i) the phenotypes studied, (ii) the data source and data types used, (iii) the ML approaches applied, and (iv) the broader context of each article (**Figure 1**). The complete list of the 52 items recorded from the articles is included in Supplementary Material **Table S5**.

RESULTS

Phenotypes

The articles in our study considered 156 unique phenotypes, with half of the articles focusing on more than one phenotype. The most common phenotypes are summarized in **Figure 3** and include chronic conditions as well as adverse drug events. Social determinants of health such as marital status and homelessness were considered in 7 of the 106 articles [25–31], while more nuanced phenotypes such as disease severity (n = 5, 4.7%) [32–36] and disease subtypes (n = 19, 17.9%) [37–54] are emerging areas of focus. We also classified each phenotype as either binary (eg. presence/absence of diabetes), categorical (eg. past/current/never smoker), or temporal (eg. progression of asthma over time, time of asthma diagnosis). The vast majority of studies focused on at least one binary phenotype (n = 84, 79.2%).

Data sources

Most studies relied on EHR data from a single institution or health system, including Partners HealthCare/Mass General Brigham (n = 7, 6.6%) and Mayo Clinic (n = 3, 2.8%). Nearly one-third of studies used openly available datasets such as datasets released by natural language processing (NLP) competitions and the Medical Information Mart for Intensive Care (MIMIC-III) [55] (Supplementary Material **Table S6**) . Several studies linked EHR data to administrative claims [39,56–61] or registry databases [45,62–64] to validate the accuracy of an EHR-based phenotyping method. Specimen biorepositories have also been used to demonstrate the accuracy of a derived phenotype in genetic applications, such as a replication of a genome-wide association study [65–68]. Recently, phenotyping algorithms have been developed in conjunction with information from publicly available knowledge bases, including PubMed, Wikipedia, Medscape [66,68–73], MedDRA and SIDER drug characterization databases [74,75], and the Unified Medical Language System (UMLS) [76,77].

Data types

The articles leveraged multiple data types for phenotyping. EHR data is available in two primary forms: (i) structured data (eg. diagnosis codes, procedure codes) and (ii) unstructured data in free-text (eg. progress notes, discharge summaries) and images[1]. A majority of articles (n = 70, 66.0%) articles used structured data as it is simpler to analyze than free-text [78,79]. Diagnosis codes were the most common form of structured data and used in 58 of the 106 articles (**Figure 4a**). We also summarize the use of standardized terminologies in **Table S7** of the Supplementary Material.

Unstructured data was used in 72 of the 106 articles, with half of the studies using both structured and unstructured data. **Figure 4b** presents the types of free-text notes used for phenotyping, though most articles did not clearly specify the note types used. In contrast to structured data, clinical notes require substantial processing due to abundant abbreviations. synonyms, clinical jargon, and misspellings [80]. Most papers used openly available software (n = 42, 39.6%) to map words into medical concepts within the UMLS Metathesaurus. Extracted medical concepts are often engineered into features for traditional ML-based phenotyping models (eg. number of positive mentions of the phenotype in a patient record) [30,33,43,81–85]. The most common clinical NLP software was clinical Text Analysis and Knowledge Extraction System (cTAKES) and was used in 20% of the articles [86]. **Table S8** in the Supplementary Material summarizes other software.

[1] We focus on free-text data for this review as there is only 1 article that utilized medical images.

An emerging trend to process clinical notes is the use of word embeddings [87,88]. A word embedding represents each word as a real-valued vector based on its context in order to group together words with similar syntactic or semantic meaning [89]. An embedding is learned from a large corpus of text without any human input (ie. feature engineering or labeling). Within the articles, corpora included notes from EHRs [33,83,90–92] as well as external sources such as biomedical publications [34,72,93–96] and articles from Wikipedia and Google News [32,69,97]. Among the methods used to train word embeddings, Word2vec [98] and Bidirectional Encoder Representations from Transformers (BERT) and variants [99–102] are the most frequently used (Supplementary Material Table S9). Word embeddings typically serve as the input layer to phenotyping algorithms using deep learning models and have also been used to account for ambiguous abbreviations and spelling errors in clinical notes [25,27,29,32,34,42,69,72,74–77,83,93–97,103–114].

Machine learning methods

The primary ML approach used in each paper was classified as either supervised, semi-supervised, weakly-supervised, unsupervised, or reinforcement learning. In reviewing the literature, we found that deep learning and federated learning appeared in the articles. Additionally, most deep learning approaches were supervised and only one paper used reinforcement learning with a deep learning approach [115]]. We therefore summarize contributions in (i) traditional supervised, (ii) deep supervised, (iii) semi-supervised, (iv) weakly-supervised, (v) unsupervised, and (vi) federated learning in the subsequent sections. Definitions for the ML approaches as well as common evaluation metrics are in Supplementary Material **Tables S10-11**.

Supervised learning

Traditional supervised learning

As shown in **Figure 5**, the majority of articles employed supervised learning models (n = 64, 60.4%), with nearly half of the articles using traditional supervised learning methods. Supervised models are trained with labeled data containing information on both the (i) features predictive of the phenotype (eg. number of billing codes and free-text mentions of the phenotype in a patient record) and (ii) the 'gold-standard' phenotype labels (eg.presence/absence of the phenotype). Traditionally, gold-standard labels are annotated by manual review of patient records [37,85,116]. Labels have also been derived from registry data [33], laboratory results [61,112,117], diagnosis codes [30,57,58,118–120], and rule-based algorithms [59,121–123] to enable more rapid development of labeled datasets. The most commonly used methods for classifying a binary phenotype are random forest [26,28,35,37,56,57,60,62,70,81,84,117,119,120,124–126], logistic regression [36,37,57,58,60,67,82,84,93,116,117,119,125,127,128], and support vector machine (SVM) [31,35,37,58,60,81,82,84,92,97,104,116,125,126] (Supplementary Material **Table S12**). It is important to note that in contrast to rule-based methods, most ML methods output a predicted score or probability of the phenotype for each patient. This provides researchers flexibility in determining a threshold to classify phenotype status for a given application.

A majority of articles (n = 18, 31%) developed accurate algorithms to phenotype chronic conditions such as cancer [62,128] and heart failure [61]. Several studies considered rare phenotypes which required methods for the imbalanced class setting [58,106]. For instance, SVM with the Synthetic Minority Oversampling Technique (SMOTE) [129] was used to identify

cases of insulin rejection, a condition documented in 0.02% of the clinical notes, with improved sensitivity relative to standard SVM (precision/recall$_{SVM+SMOTE}$ 0.79/0.70, precision/recall$_{SVM}$ 0.90/0.58) [130]. However, it has been noted that methods for the imbalance setting do not uniformly yield improved performance [60,91]. Seneviratne et al. observed that precision and recall dropped substantially when adding SMOTE to a random forest model classifying metastatic prostate cancer cases [62].

Deep supervised learning

A limitation of traditional supervised learning methods is their inability to handle raw input data, which requires researchers to manually curate features for model training [131]. To alleviate the burden of feature engineering, deep learning models consist of many processing layers that discover intrinsic patterns in data [132]. 33 of the 72 supervised learning articles employ deep learning, with 31 of these articles using word embeddings as the input layer. Feed-forward neural networks are the least frequently used architecture [25,32,108,111,118,133], while recurrent neural networks (RNNS), [27,34,59,69,74–76,85,90,94,95,103,105–107,109,110,123,126,130,134], convolutional neural networks (CNNs) [33,75,83,91,93,94,103,106,107,112,126,134–136], and their variants are frequently used (Supplementary Material **Table S12**). RNNs are particularly well suited for sequential data from longitudinal EHRs. Within the selected articles, RNNs and their variants such as long-short term memory (LSTM) networks [137] have been applied to characterize diseases with symptoms evolving over time with high performance, such as sepsis (AUROC$_{RNN}$ 0.88) [121] and pediatric asthma (F1-score$_{LSTM}$ 0.85) [123]. Gao et al. compared the performance of a reinforcement learning RNN model with logistic regression for phenotyping 25

chronic or critical conditions and demonstrated that RNNs enhanced performance by up to 7% across phenotypes (Macro-AUROC$_{\text{Logistic regression}}$ 0.73, Macro-AUROC$_{\text{RNN}}$ 0.78) [115].

CNNs are commonly used for classification tasks with free-text data. In the context of phenotyping, CNNs have been applied to identify conditions primarily documented in clinical notes with high performance, including anaphylaxis (F1-score 0.96) and substance abuse disorders such as alcohol, drug, and tobacco abuse (F1-score 0.88-0.95) [103]. Gehrmann et al. demonstrated that CNNs outperformed logistic regression in phenotyping obesity, depression, chronic pain, and 7 other phenotypes using the MIMIC-III phenotype annotation dataset (F1-score$_{\text{CNN}}$ 0.57-0.97, F1-score$_{\text{logistic}}$ 0.39-0.82) [83].

In spite of the advantages of deep learning, traditional supervised learning methods have outperformed their deep learning counterparts. For example, Martin et al. found that a penalized elastic net logistic regression consistently outperformed FFNNs for identifying frailty aspects such as musculoskeletal problems and respiratory impairment (F1-score$_{\text{Elastic net}}$ 0.51-0.65, F1-score$_{\text{FFNN}}$ 0.24-0.62)[95]. Several authors also commented on challenges related to model interpretability and transportability of deep learning models across sites [25,85,91,130].

Semi-supervised learning

Although supervised learning is the most common paradigm for ML-based phenotyping, obtaining high-quality labeled data through chart review is a resource intensive task [138]. Semi-supervised learning methods are trained with a large amount of unlabeled data (ie. unreview medical records), together with a small amount of labeled data, in order to minimize

the burden of chart review [139]. 5 articles used semi-supervised learning methods. Estiri et al. utilized a self-learning approach to develop standard generative models (eg. Naive Bayes, Linear Discriminant Analysis) using a small set of labeled data (average 182 patients) and larger set of unlabeled (average 5956 patients) for classification of 18 phenotypes [114]. The approach performed on par with supervised learning, but required less labeled data (AUROC 0.78-0.99). Estiri et al. later combined this method with a sequential pattern mining approach to capture temporal information (eg. sequence of diagnoses) that is underutilized in phenotyping [113]. The model achieved superior performance over previous studies utilizing supervised learning across 5 time-evolving phenotypes (eg.$AUROC_{semi\text{-}supervised}$ 0.98, $AUROC_{superivsed}$ 0.90-0.93 for congestive heart failure) [113]. Zhang et al. developed a logistic regression method using positive-only data with 150 cases identified by chart review and 6000 unlabeled records for phenotyping primary aldosteronism (recall/precision 0.85/0.78) [140]. A similar approach using SVM was applied to identify cases of acute hepatic porphyria (AUROC 0.775) [141]. Henderson et al. developed a semi-supervised tensor factorization method that underperformed relative to supervised learning in phenotyping resistant hypertension and type-2 diabetes, but required a fraction of reviewed charts (e.g $AUROC_{semi\text{-}supervised}$ 0.66, $AUROC_{superivsed}$ 0.69 for hypertension) [142].

Weakly-supervised learning

To eliminate the need for gold-standard labeled data, an increasingly common ML approach for phenotyping is weakly-supervised learning. The key idea behind weakly-supervised learning is to utilize a “silver-standard” label that can be easily extracted from all records in place of the gold-standard obtained from chart review [143]. The silver-standard label is typically selected

by domain experts based on clinical expertise as a variable that is a highly predictive, but imperfect proxy for the gold-standard [143–146]. For instance, in a study of systemic lupus erythematosus, the silver standard labeled patients as having the condition if 4 or more disease-specific ICD-9 codes were present in their record and as a control otherwise [147]. Other common choices of silver standard labels include lab results and free-text mentions of the phenotype of interest [66,68,71], or topic categories defined by word embeddings [63].

Two main approaches for weakly-supervised learning appeared in the articles. The first assumes that the silver-standard label follows a mixture model representing phenotype cases and controls [65,66,68,71–73,148]. For example, PheNorm [71] utilized Gaussian mixture-modeling and denoising self-regression with silver standard labels based on counts of relevant billing codes and free-text mentions of the phenotype. PheNorm achieved comparable accuracy to penalized logistic regression trained with 100-300 labels for 4 phenotypes ($AUROC_{PheNorm}$ 0.90-0.95, $AUC_{Logistic}$ 0.90-0.95) [71]. The Multimodal Automated Phenotyping (MAP) method later improved upon PheNorm by using an ensemble of mixture models and outperformed PheCodes for 16 phenotypes [64,66]. The second set of approaches uses supervised learning methods trained with a silver standard derived from data elements whose presence indicates a patient has the phenotype whereas the absence is uninformative or is a noisy measure of control status [25,77,96,144,145,147,149,150]. For example, the Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE) based on silver-standard logistic regression was applied to 10 phenotypes across 3 Observational Health Data Sciences and Informatics (OHDSI) sites in the U.S. and Korea [149]. APHRODITE achieved mean recall and precision of 0.54 and 0.73 compared to 0.46 and 0.24 in Korea. The

decreased performance at the Korean site is likely due to regional differences in the quality of silver-standard labels. Similar approaches utilizing deep learning approaches have been developed in [25].

Unsupervised learning

In contrast to the previously mentioned approaches, unsupervised learning is primarily used for phenotype discovery [151]. 20 articles developed unsupervised models, with a majority using latent Dirichlet allocation (LDA) [38,41,43,46,48], matrix or tensor factorization [40,44,52], and hierarchical clustering [45,47,49–51] (Supplementary Material **Table S12**). A deep autoencoder was used to discover depression and suicidal ideation patterns [47]. Discovering subphenotypes [39,43,46,48–50] and clusters of disease trajectory patterns [41,45,47,53,54] are the main focus of unsupervised models. For example, non-negative tensor factorization was used to discover 14 subphenotypes of cardiovascular disease (CVD) patients and their association with conventional CVD risk scores and subsequent myocardial infarction [44]. Ibrahim et al. used hierarchical clustering to identify sepsis clusters of patients with distinct organ dysfunction patterns [50].

Federated learning

With progress towards interoperability, such as common data models (CDMs) [151–154] and Fast Healthcare Interoperability Resources (FHIR) [125,155], there has been increased interest in developing phenotyping algorithms using EHR data from multiple institutions. Federated learning is a ML paradigm that facilitates model training with data from multiple institutions utilizing only summary information (eg. model parameter estimate) in order to protect against the sharing of individual-level data (Supplementary Material Figure S1). Choudhury et al. extended

federated learning to SVM, single-layer perceptron, and logistic regression [58]. The federated approaches generally improved the precision for identifying opioid prescriptions relative to localized training at a single institution ($Precision_{Federated}$ 0.92, $Precision_{Localized}$ 0.95). Kuo et al. utilized blockchain technology for the classification of myocardial infarction with similar performance as centralized learning using aggregated individual-level data ($AUC_{blockchain}$ 0.954-0.965, $AUC_{centralized}$ 0.957-0.965) [128].

Context

Primary contributions

Figure 6 presents a summary of papers by their primary contribution. Overall, application papers are the most common type and primarily utilize supervised learning methods. Applications developed phenotyping algorithms for diverse uses, including development of clinical decision support tools [51,93,95] and phenome-wide association studies [66,68]. Among methods articles, approaches to alleviate the burden of labeling, such as semi-supervised and weakly-supervised learning, are the most prevalent.

Evaluation and reporting methods

Table 1 presents the evaluation metrics articles used to evaluate the performance of phenotyping algorithms. Because phenotyping is typically focused on accurately identifying phenotype cases, precision, recall, and F-score are most commonly reported. However, there are inconsistencies in reporting of multiple metrics across the articles and model calibration is rarely assessed. Additionally, ML models were only compared with rule-based methods in 22 articles [25,26,36,37,60,62–66,68,81,82,85,92,104,107,116,121,125,130,150] and exhibited improved

performance in all but 4 studies [25,26,130,150]. Overall, we found limited reporting of data descriptors necessary for implementation of a phenotyping algorithm to a new setting. For instance, patient demographics were only reported in 43 or 74 of papers using private data sources. Additionally, only 21% of articles released code. One third of methodology papers, which use non-standard approaches unavailable in analytic software, were accompanied by open source code.

DISCUSSION

Deep phenotyping

In recent years, there has been a call for "deep phenotyping" methods that move beyond identification of binary phenotypes to characterization of more complex phenotypes such as timing or severity of a condition [24,44–46,115,156]. While our study indicates that existing literature remains focused on characterizing binary phenotypes, several preprints consider severity and temporal phenotyping [36,157,158]. Additionally, most articles relied solely on EHR data. Characterizing nuanced phenotypes will require the integration of diverse datasets, including voice recordings, genomic data, and mobile and wearable data, that provide the behavioral, genetic, and environmental information that is missing from EHRs.

Evaluation & Reporting standards

Most articles used standard metrics to evaluate algorithm performance. It is becoming increasingly apparent that ML models are prone to bias that can harm marginalized groups [159,160]. Only 1 article evaluated algorithmic fairness [161]. Fairness must be integrated into phenotyping in the future. However, reliable evaluation of model performance requires a

substantial amount of gold-standard labeled data which is cumbersome to obtain. There is very little work on semi-supervised and weakly-supervised approaches for evaluating model performance and further work is warranted [162]. Moreover, with the added complexity of ML approaches and the increased use of EHR for health research, there is an urgent need for reproducible phenotype definitions to enhance replicability and quality. Kohane et al. recently introduced a checklist for evaluating the quality of EHR-based studies, emphasizing phenotypic transparency as a key concern [164]. OHDSI has proposed FAIR (Findable, Accessible, Interoperable, Reusable) phenotype definitions [163], but they have yet to be widely adopted.

Accounting for phenotyping error

As ML expands the scope of phenotyping, care must be taken when using derived phenotypes for downstream tasks as they are subject to error. In the context of association studies, it is well known within the statistical community that imperfect phenotypes lead to diminished statistical power and biased estimation [165–167]. However, statistical methods are often siloed from the informatics community. We anticipate more work in disseminating existing methods as well as in developing methodology for more complex settings such as predictive modeling studies.

Limitations

As the definition of phenotyping is broad and varies within the literature [11], we used a search capturing articles focusing on ML or phenotyping using EHRs. Following prior studies, we limited our scope to journals and conferences with the highest expected publications of recent phenotyping methods [14]. Although we have missed articles published outside of these

journals, our primary focus was to characterize key trends in ML-based phenotyping, which we believe are captured in the venues considered in our study.

CONCLUSION

This review provides a summary of the recent trends of ML-based phenotyping between 2018 and 2022. While traditional supervised learning is the most prevalent, semi-supervised, weakly-supervised, deep, and federated learning are expanding the scope of phenotyping across conditions and healthcare settings. Current literature has laid the groundwork for "deep phenotyping", but developing standards and methodology for reliable use of phenotypes derived from ML models is necessary for the continued success of EHR-based research.

REFERENCES

1. Institute of Medicine, Roundtable on Value and Science-Driven Health Care. Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary. National Academies Press; 2011.

2. Liao KP, Diogo D, Cui J, Cai T, Okada Y, Gainer VS, et al. Association between low density lipoprotein and rheumatoid arthritis genetic factors with low density lipoprotein levels in rheumatoid arthritis and non-rheumatoid arthritis controls. Ann Rheum Dis. 2014;73: 1170–1175.

3. Xia Z, Secor E, Chibnik LB, Bove RM, Cheng S, Chitnis T, et al. Modeling disease severity in multiple sclerosis using electronic health records. PLoS One. 2013;8: e78927.

4. Mc Cord KA, Hemkens LG. Using electronic health records for clinical trials: Where do we stand and where can we go? CMAJ. 2019;191: E128–E133.

5. Kim JH, Ta CN, Liu C, Sung C, Butler AM, Stewart LA, et al. Towards clinical data-driven eligibility criteria optimization for interventional COVID-19 clinical trials. J Am Med Inform Assoc. 2021;28: 14–22.

6. Abul-Husn NS, Kenny EE. Personalized Medicine and the Power of Electronic Health Records. Cell. 2019;177: 58–69.

7. Franklin JM, Liaw K-L, Iyasu S, Critchlow CW, Dreyer NA. Real-world evidence to support regulatory decision making: New or expanded medical product indications. Pharmacoepidemiol Drug Saf. 2021;30: 685–693.

8. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015;350: h1885.

9. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Medicine. 2015. doi:10.1186/s13073-015-0166-y

10. Pendergrass SA, Crawford DC. Using Electronic Health Records To Generate Phenotypes For Research. Curr Protoc Hum Genet. 2019;100: e80.

11. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014;21: 221–230.

12. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating EHR phenotypes: CALIBER. bioRxiv. 2019. p. 539403. doi:10.1101/539403

13. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of

electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 2013;20: e147–54.

14. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci. 2018;1: 53–68.

15. Kern EFO, Maney M, Miller DR, Tseng C-L, Tiwari A, Rajan M, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. Health Serv Res. 2006;41: 564–580.

16. Wei W-Q, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc. 2012;19: 219–224.

17. Alzoubi H, Alzubi R, Ramzan N, West D, Al-Hadhrami T, Alazab M. A Review of Automatic Phenotyping Approaches using Electronic Health Records. Electronics. 2019;8: 1235.

18. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc. 2018;25: 1419–1428.

19. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural Language Processing for EHR-Based Computational Phenotyping. IEEE/ACM Trans Comput Biol Bioinform. 2019;16: 139–153.

20. Robinson JR, Wei W-Q, Roden DM, Denny JC. Defining Phenotypes from Clinical Data to Drive Genomic Research. Annu Rev Biomed Data Sci. 2018;1: 69–92.

21. McBrien KA, Souri S, Symonds NE, Rouhi A, Lethebe BC, Williamson TS, et al. Identification of validated case definitions for medical conditions used in primary care electronic medical record databases: a systematic review. J Am Med Inform Assoc. 2018;25: 1567–1578.

22. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc. 2016;23: 1007–1015.

23. Irwin AN, Rackham D. Comparison of the time-to-indexing in PubMed between biomedical journals according to impact factor, discipline, and focus. Res Social Adm Pharm. 2017;13: 389–393.

24. Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. J Am Med Inform Assoc. 2018;25: 289–294.

25. Annapragada AV, Donaruma-Kwoh MM, Annapragada AV, Starosolski ZA. A natural language processing and deep learning approach to identify child abuse from pediatric

electronic medical records. PLoS One. 2021;16: e0247404.

26. Bucher BT, Shi J, Pettit RJ, Ferraro J, Chapman WW, Gundlapalli A. Determination of Marital Status of Patients from Structured and Unstructured Electronic Healthcare Data. AMIA Annu Symp Proc. 2019;2019: 267–274.

27. Stemerman R, Arguello J, Brice J, Krishnamurthy A, Houston M, Kitzmiller R. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. JAMIA Open. 2021;4: ooaa069.

28. Carson NJ, Mullin B, Sanchez MJ, Lu F, Yang K, Menezes M, et al. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. PLoS One. 2019;14: e0211116.

29. Bejan CA, Angiolillo J, Conway D, Nash R, Shirey-Rice JK, Lipworth L, et al. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. Journal of the American Medical Informatics Association. 2018. pp. 61–71. doi:10.1093/jamia/ocx059

30. Buckland RS, Hogan JW, Chen ES. Selection of Clinical Text Features for Classifying Suicide Attempts. AMIA Annu Symp Proc. 2020;2020: 273–282.

31. Feller DJ, Zucker J, Don't Walk OB, Srikishan B, Martinez R, Evans H, et al. Towards the Inference of Social and Behavioral Determinants of Sexual Health: Development of a Gold-Standard Corpus with Semi-Supervised Learning. AMIA Annu Symp Proc. 2018;2018: 422–429.

32. Mitra A, Rawat BPS, McManus D, Kapoor A, Yu H. Bleeding entity recognition in electronic health records: A comprehensive analysis of end-to-end systems. AMIA Annu Symp Proc. 2020;2020: 860–869.

33. Kulshrestha S, Dligach D, Joyce C, Gonzalez R, O'Rourke AP, Glazer JM, et al. Comparison and interpretability of machine learning models to predict severity of chest injury. JAMIA Open. 2021. doi:10.1093/jamiaopen/ooab015

34. Lybarger K, Ostendorf M, Thompson M, Yetisgen M. Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. J Biomed Inform. 2021;117: 103761.

35. Badger J, LaRose E, Mayer J, Bashiri F, Page D, Peissig P. Machine learning for phenotyping opioid overdose events. J Biomed Inform. 2019;94: 103185.

36. Klann JG, Estiri H, Weber GM, Moal B, Avillach P, Hong C, et al. Validation of an internationally derived patient severity phenotype to support COVID-19 analytics from electronic health record data. J Am Med Inform Assoc. 2021;28: 1411–1420.

37. Ni Y, Alwell K, Moomaw CJ, Woo D, Adeoye O, Flaherty ML, et al. Towards phenotyping

stroke: Leveraging data from a large-scale epidemiological study to detect stroke diagnosis. PLoS One. 2018;13: e0192586.

38. Bhattacharya M, Jurkovitz C, Shatkay H. Co-occurrence of medical conditions: Exposing patterns through probabilistic topic modeling of snomed codes. J Biomed Inform. 2018;82: 31–40.

39. Prenovost KM, Fihn SD, Maciejewski ML, Nelson K, Vijan S, Rosland A-M. Using item response theory with health system data to identify latent groups of patients with multiple health conditions. PLoS One. 2018;13: e0206915.

40. Perros I, Papalexakis EE, Vuduc R, Searles E, Sun J. Temporal phenotyping of medically complex children via PARAFAC2 tensor factorization. J Biomed Inform. 2019;93: 103125.

41. Wang L, Lakin J, Riley C, Korach Z, Frain LN, Zhou L. Disease Trajectories and End-of-Life Care for Dementias: Latent Topic Modeling and Trend Analysis Using Clinical Notes. AMIA Annu Symp Proc. 2018;2018: 1056–1065.

42. Shen F, Peng S, Fan Y, Wen A, Liu S, Wang Y, et al. HPO2Vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the Human Phenotype Ontology. J Biomed Inform. 2019;96: 103246.

43. Afshar M, Joyce C, Dligach D, Sharma B, Kania R, Xie M, et al. Subtypes in patients with opioid misuse: A prognostic enrichment strategy using electronic health record data in hospitalized patients. PLoS One. 2019;14: e0219717.

44. Zhao J, Zhang Y, Schlueter DJ, Wu P, Eric Kerchberger V, Trent Rosenbloom S, et al. Detecting time-evolving phenotypic topics via tensor factorization on electronic health records: Cardiovascular disease case study. J Biomed Inform. 2019;98: 103270.

45. Lyudovyk O, Shen Y, Tatonetti NP, Hsiao SJ, Mansukhani MM, Weng C. Pathway analysis of genomic pathology tests for prognostic cancer subtyping. J Biomed Inform. 2019;98: 103286.

46. Liu Q, Woo M, Zou X, Champaneria A, Lau C, Mubbashar MI, et al. Symptom-based patient stratification in mental illness using clinical notes. J Biomed Inform. 2019;98: 103274.

47. Gong J, Simon GE, Liu S. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. PLoS One. 2019;14: e0222665.

48. Wang Y, Zhao Y, Therneau TM, Atkinson EJ, Tafti AP, Zhang N, et al. Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. Journal of Biomedical Informatics. 2020. p. 103364. doi:10.1016/j.jbi.2019.103364

49. Xu Z, Chou J, Zhang XS, Luo Y, Isakova T, Adekkanattu P, et al. Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health

record data with memory networks. J Biomed Inform. 2020;102: 103361.

50. Ibrahim ZM, Wu H, Hamoud A, Stappen L, Dobson RJB, Agarossi A. On classifying sepsis heterogeneity in the ICU: insight using machine learning. J Am Med Inform Assoc. 2020;27: 437–443.

51. Apostolova E, Uppal A, Galarraga JE, Koutroulis I, Tschampel T, Wang T, et al. Towards reliable ARDS clinical decision support: ARDS patient analytics with free-text and structured EMR data. AMIA Annu Symp Proc. 2019;2019: 228–237.

52. Hassaine A, Canoy D, Solares JRA, Zhu Y, Rao S, Li Y, et al. Learning multimorbidity patterns from electronic health records using Non-negative Matrix Factorisation. J Biomed Inform. 2020;112: 103606.

53. Hubbard RA, Xu J, Siegel R, Chen Y, Eneli I. Studying pediatric health outcomes with electronic health records using Bayesian clustering and trajectory analysis. J Biomed Inform. 2021;113: 103654.

54. Mullin S, Zola J, Lee R, Hu J, MacKenzie B, Brickman A, et al. Longitudinal K-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes. J Biomed Inform. 10/2021;122: 103889.

55. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3: 160035.

56. Erickson J, Abbott K, Susienka L. Automatic address validation and health record review to identify homeless Social Security disability applicants. J Biomed Inform. 2018;82: 41–46.

57. Fialoke S, Malarstig A, Miller MR, Dumitriu A. Application of Machine Learning Methods to Predict Non-Alcoholic Steatohepatitis (NASH) in Non-Alcoholic Fatty Liver (NAFL) Patients. AMIA Annu Symp Proc. 2018;2018: 430–439.

58. Choudhury O, Park Y, Salonidis T, Gkoulalas-Divanis A, Sylla I, Das AK. Predicting Adverse Drug Reactions on Distributed Health Data using Federated Learning. AMIA Annu Symp Proc. 2019;2019: 313–322.

59. Nori VS, Hane CA, Sun Y, Crown WH, Bleicher PA. Deep neural network models for identifying incident dementia using claims and EHR datasets. PLoS One. 2020;15: e0236400.

60. Gibson TB, Nguyen MD, Burrell T, Yoon F, Wong J, Dharmarajan S, et al. Electronic phenotyping of health outcomes of interest using a linked claims-electronic health record database: Findings from a machine learning pilot project. J Am Med Inform Assoc. 2021;28: 1507–1517.

61. Mahesri M, Chin K, Kumar A, Barve A, Studer R, Lahoz R, et al. External validation of a claims-based model to predict left ventricular ejection fraction class in patients with heart failure. PLoS One. 2021;16: e0252903.

62. Seneviratne MG, Banda JM, Brooks JD, Shah NH, Hernandez-Boussard TM. Identifying Cases of Metastatic Prostate Cancer Using Machine Learning on Electronic Health Records. AMIA Annu Symp Proc. 2018;2018: 1498–1504.

63. Ling AY, Kurian AW, Caswell-Jin JL, Sledge GW Jr, Shah NH, Tamang SR. Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. JAMIA Open. 2019;2: 528–537.

64. Geva A, Liu M, Panickan VA, Avillach P, Cai T, Mandl KD. A high-throughput phenotyping algorithm is portable from adult to pediatric populations. J Am Med Inform Assoc. 2021;28: 1265–1269.

65. Sinnott JA, Cai F, Yu S, Hejblum BP, Hong C, Kohane IS, et al. PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies. J Am Med Inform Assoc. 2018;25: 1359–1365.

66. Liao KP, Sun J, Cai TA, Link N, Hong C, Huang J, et al. High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. J Am Med Inform Assoc. 2019;26: 1255–1262.

67. Li R, Chen Y, Moore JH. Integration of genetic and clinical information to improve imputation of data missing from electronic health records. J Am Med Inform Assoc. 2019;26: 1056–1063.

68. Zheng NS, Feng Q, Kerchberger VE, Zhao J, Edwards TL, Cox NJ, et al. PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. J Am Med Inform Assoc. 2020;27: 1675–1687.

69. Si Y, Roberts K. A Frame-Based NLP System for Cancer-Related Information Extraction. AMIA Annu Symp Proc. 2018;2018: 1524–1533.

70. Goodwin TR, Harabagiu SM. Learning relevance models for patient cohort retrieval. JAMIA Open. 2018;1: 265–275.

71. Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, Gainer VS, et al. Enabling phenotypic big data with PheNorm. J Am Med Inform Assoc. 2018;25: 54–60.

72. Ning W, Chan S, Beam A, Yu M, Geva A, Liao K, et al. Feature extraction for phenotyping from semantic and knowledge resources. J Biomed Inform. 2019;91: 103122.

73. Ferté T, Cossin S, Schaeverbeke T, Barnetche T, Jouhet V, Hejblum BP. Automatic phenotyping of electronical health record: PheVis algorithm. J Biomed Inform. 2021;117: 103746.

74. Ju M, Nguyen NTH, Miwa M, Ananiadou S. An ensemble of neural models for nested adverse drug events and medication extraction with subwords. J Am Med Inform Assoc. 2020;27: 22–30.

75. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. J Am Med Inform Assoc. 2020;27: 65–72.

76. Chen L, Gu Y, Ji X, Sun Z, Li H, Gao Y, et al. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. J Am Med Inform Assoc. 2020;27: 56–64.

77. Banerjee I, Li K, Seneviratne M, Ferrari M, Seto T, Brooks JD, et al. Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. JAMIA Open. 2019. pp. 150–159. doi:10.1093/jamiaopen/ooy057

78. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. Artif Intell Med. 2016;71: 57–61.

79. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12: 417–428.

80. Agaronnik ND, Lindvall C, El-Jawahri A, He W, Iezzoni LI. Challenges of Developing a Natural Language Processing Method With Electronic Health Records to Identify Persons With Chronic Mobility Disability. Arch Phys Med Rehabil. 2020;101: 1739–1746.

81. Koola JD, Davis SE, Al-Nimri O, Parr SK, Fabbri D, Malin BA, et al. Development of an automated phenotyping algorithm for hepatorenal syndrome. J Biomed Inform. 2018;80: 87–95.

82. Afshar M, Phillips A, Karnik N, Mueller J, To D, Gonzalez R, et al. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. J Am Med Inform Assoc. 2019;26: 254–261.

83. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One. 2018;13: e0192360.

84. Shi J, Liu S, Pruitt LCC, Luppens CL, Ferraro JP, Gundlapalli AV, et al. Using Natural Language Processing to improve EHR Structured Data-based Surgical Site Infection Surveillance. AMIA Annu Symp Proc. 2019;2019: 794–803.

85. Ni Y, Bachtel A, Nause K, Beal S. Automated detection of substance use information from electronic health records for a pediatric population. J Am Med Inform Assoc. 2021;28: 2116–2127.

86. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17: 507–513.

87. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. Journal of Biomedical Informatics: X. 2019;4: 100057.

88. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc. 2020;27: 457–470.

89. Teller V. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …; 2000. Available: https://direct.mit.edu/coli/article-abstract/26/4/638/1680

90. Guan M, Cho S, Petro R, Zhang W, Pasche B, Topaloglu U. Natural language processing and recurrent network models for identifying genomic mutation-associated cancer treatment change from patient progress notes. JAMIA Open. 2019;2: 139–149.

91. Segura-Bedmar I, Colón-Ruíz C, Tejedor-Alonso MÁ, Moro-Moro M. Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. J Biomed Inform. 2018;87: 50–59.

92. Dai H-J, Wang F-D, Chen C-W, Su C-H, Wu C-S, Jonnagaddala J. Cohort selection for clinical trials using multiple instance learning. J Biomed Inform. 2020;107: 103438.

93. Obeid JS, Davis M, Turner M, Meystre SM, Heider PM, O'Bryan EC, et al. An artificial intelligence approach to COVID-19 infection risk assessment in virtual visits: A case report. J Am Med Inform Assoc. 2020;27: 1321–1325.

94. Dai H-J, Su C-H, Wu C-S. Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. Journal of the American Medical Informatics Association. 2020. pp. 47–55. doi:10.1093/jamia/ocz120

95. Martin JA, Crane-Droesch A, Lapite FC, Puhl JC, Kmiec TE, Silvestri JA, et al. Development and validation of a prediction model for actionable aspects of frailty in the text of clinicians' encounter notes. J Am Med Inform Assoc. 2021;29: 109–119.

96. Sanyal J, Rubin D, Banerjee I. A weakly supervised model for the automated detection of adverse events using clinical notes. J Biomed Inform. 2022;126: 103969.

97. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform. 2018;87: 12–20.

98. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2013. Available: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

99. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv [cs.CL]. 2018. Available: http://arxiv.org/abs/1810.04805

100. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36: 1234–1240.

101. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. arXiv [cs.CL]. 2019. Available: http://arxiv.org/abs/1904.03323

102. Liu Z, Lin W, Shi Y, Zhao J. A Robustly Optimized BERT Pre-training Approach with Post-training. Lecture Notes in Computer Science. 2021. pp. 471–484. doi:10.1007/978-3-030-84186-7_31

103. Lybarger K, Yetisgen M, Ostendorf M. Using Neural Multi-task Learning to Extract Substance Abuse Information from Clinical Notes. AMIA Annu Symp Proc. 2018;2018: 1395–1404.

104. Liu F, Pradhan R, Druhl E, Freund E, Liu W, Sauer BC, et al. Learning to detect and understand drug discontinuation events from clinical narratives. J Am Med Inform Assoc. 2019;26: 943–951.

105. Chen T, Dredze M, Weiner JP, Kharrazi H. Identifying vulnerable older adult populations by contextualizing geriatric syndrome information in clinical notes of electronic health records. J Am Med Inform Assoc. 2019;26: 787–795.

106. Xiong Y, Shi X, Chen S, Jiang D, Tang B, Wang X, et al. Cohort selection for clinical trials using hierarchical neural network. J Am Med Inform Assoc. 2019;26: 1203–1208.

107. Ji B, Li S, Yu J, Ma J, Tang J, Wu Q, et al. Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models. J Biomed Inform. 2020;104: 103395.

108. Soni S, Roberts K. Patient Cohort Retrieval using Transformer Language Models. AMIA Annu Symp Proc. 2020;2020: 1150–1159.

109. Wei Q, Ji Z, Li Z, Du J, Wang J, Xu J, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. J Am Med Inform Assoc. 2019;27: 13–21.

110. Kim Y, Meystre SM. Ensemble method-based extraction of medication and related information from clinical texts. J Am Med Inform Assoc. 2020;27: 31–38.

111. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. J Am Med Inform Assoc. 2020;27: 1935–1942.

112. Meystre SM, Heider PM, Kim Y, Davis M, Obeid J, Madory J, et al. Natural language processing enabling COVID-19 predictive analytics to support data-driven patient advising and pooled testing. J Am Med Inform Assoc. 2021;29: 12–21.

113. Estiri H, Strasser ZH, Murphy SN. High-throughput phenotyping with temporal sequences. J Am Med Inform Assoc. 2020;28: 772–781.

114. Estiri H, Vasey S, Murphy SN. Generative transfer learning for measuring plausibility of EHR diagnosis records. J Am Med Inform Assoc. 2021;28: 559–568.

115. Gao J, Xiao C, Glass LM, Sun J. Dr. Agent: Clinical predictive model via mimicked second opinions. J Am Med Inform Assoc. 2020;27: 1084–1091.

116. Afshar M, Joyce C, Oakey A, Formanek P, Yang P, Churpek MM, et al. A Computable Phenotype for Acute Respiratory Distress Syndrome Using Natural Language Processing and Machine Learning. AMIA Annu Symp Proc. 2018;2018: 157–165.

117. Inaguma D, Kitagawa A, Yanagiya R, Koseki A, Iwamori T, Kudo M, et al. Increasing tendency of urine protein is a risk factor for rapid eGFR decline in patients with CKD: A machine learning-based prediction model by using a big database. PLoS One. 2020;15: e0239262.

118. Ogunyemi OI, Gandhi M, Lee M, Teklehaimanot S, Daskivich LP, Hindman D, et al. Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system. JAMIA Open. 2021;4: ooab066.

119. Docherty M, Regnier SA, Capkun G, Balp M-M, Ye Q, Janssens N, et al. Development of a novel machine learning model to predict presence of nonalcoholic steatohepatitis. J Am Med Inform Assoc. 2021;28: 1235–1241.

120. Cava WL, Bauer C, Moore JH, Pendergrass SA. Interpretation of machine learning predictions for patient outcomes in electronic health records. AMIA Annu Symp Proc. 2019;2019: 572–581.

121. Bedoya AD, Futoma J, Clement ME, Corey K, Brajer N, Lin A, et al. Machine learning for early detection of sepsis: an internal and temporal validation study. JAMIA Open. 2020;3: 252–260.

122. Yang X, Gong Y, Waheed N, March K, Hogan WR, Wu Y. Identifying Cancer Patients at Risk for Heart Failure Using Machine Learning Methods. : 9.

123. Wu S, Liu S, Sohn S, Moon S, Wi C-I, Juhn Y, et al. Modeling asynchronous event sequences with RNNs. J Biomed Inform. 2018;83: 167–177.

124. Grundmeier RW, Xiao R, Ross RK, Ramos MJ, Karavite DJ, Michel JJ, et al. Identifying surgical site infections in electronic health data using predictive models. J Am Med Inform Assoc. 2018;25: 1160–1166.

125. Hong N, Wen A, Stone DJ, Tsuji S, Kingsbury PR, Rasmussen LV, et al. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. Journal of Biomedical Informatics. 2019. p. 103310. doi:10.1016/j.jbi.2019.103310

126. Tang F, Xiao C, Wang F, Zhou J. Predictive modeling in urgent care: a comparative study of machine learning approaches. JAMIA Open. 2018;1: 87–98.

127. Kreuger AL, Middelburg RA, Beckers EAM, de Vooght KMK, Zwaginga JJ, Kerkhoffs J-LH, et al. The identification of cases of major hemorrhage during hospitalization in patients with acute leukemia using routinely recorded healthcare data. PLoS One. 2018;13: e0200655.

128. Kuo T-T, Gabriel RA, Cidambi KR, Ohno-Machado L. EXpectation Propagation LOgistic REgRession on permissioned blockCHAIN (ExplorerChain): decentralized online healthcare/genomics predictive model learning. J Am Med Inform Assoc. 2020;27: 747–756.

129. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16: 321–357.

130. Malmasi S, Ge W, Hosomura N, Turchin A. Comparing information extraction techniques for low-prevalence concepts: The case of insulin rejection by patients. J Biomed Inform. 2019;99: 103306.

131. Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. 2014 Science and Information Conference. 2014. doi:10.1109/sai.2014.6918213

132. Yan LC, Yoshua B, Geoffrey H. Deep learning. Nature. 2015;521: 436–444.

133. Eisman AS, Shah NR, Eickhoff C, Zerveas G, Chen ES, Wu W-C, et al. Extracting Angina Symptoms from Clinical Notes Using Pre-Trained Transformer Architectures. AMIA Annu Symp Proc. 2020;2020: 412–421.

134. Chu J, Dong W, He K, Duan H, Huang Z. Using neural attention networks to detect adverse medical events from electronic health records. J Biomed Inform. 2018;87: 118–130.

135. Chen C-J, Warikoo N, Chang Y-C, Chen J-H, Hsu W-L. Medical knowledge infused convolutional neural networks for cohort selection in clinical trials. J Am Med Inform Assoc. 2019;26: 1227–1236.

136. MacLean MT, Jehangir Q, Vujkovic M, Ko Y-A, Litt H, Borthakur A, et al. Quantification of abdominal fat from computed tomography using deep learning and its association with electronic health records in an academic biobank. J Am Med Inform Assoc. 2021;28: 1178–1187.

137. Yu Y, Si X, Hu C, Zhang J. A Review of Recurrent Neural Networks: LSTM Cells and

Network Architectures. Neural Comput. 2019;31: 1235–1270.

138. Cai T, Cai F, Dahal KP, Cremone G, Lam E, Golnik C, et al. Improving the Efficiency of Clinical Trial Recruitment Using an Ensemble Machine Learning to Assist With Eligibility Screening. ACR Open Rheumatol. 2021;3: 593–600.

139. Zhu X (jerry). Semi-supervised learning literature survey. 2008 [cited 19 Apr 2022]. Available: https://minds.wisconsin.edu/handle/1793/60444

140. Zhang L, Ding X, Ma Y, Muthu N, Ajmal I, Moore JH, et al. A maximum likelihood approach to electronic health record phenotyping using positive and unlabeled patients. J Am Med Inform Assoc. 2020;27: 119–126.

141. Cohen AM, Chamberlin S, Deloughery T, Nguyen M, Bedrick S, Meninger S, et al. Detecting rare diseases in electronic health records using machine learning and knowledge engineering: Case study of acute hepatic porphyria. PLoS One. 2020;15: e0235574.

142. Henderson J, He H, Malin BA, Denny JC, Kho AN, Ghosh J, et al. Phenotyping through Semi-Supervised Tensor Factorization (PSST). AMIA Annu Symp Proc. 2018;2018: 564–573.

143. Yu S, Chakrabortty A, Liao KP, Cai T, Ananthakrishnan AN, Gainer VS, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. J Am Med Inform Assoc. 2017;24: e143–e149.

144. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc. 2016;23: 731–740.

145. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc. 2016;23: 1166–1173.

146. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc. 2017;2017: 48–57.

147. Murray SG, Avati A, Schmajuk G, Yazdany J. Automated and flexible identification of complex disease: building a model for systemic lupus erythematosus using noisy labeling. Journal of the American Medical Informatics Association. 2019. pp. 61–65. doi:10.1093/jamia/ocy154

148. Ahuja Y, Zhou D, He Z, Sun J, Castro VM, Gainer V, et al. sureLDA: A multidisease automated phenotyping method for the electronic health record. J Am Med Inform Assoc. 2020;27: 1235–1243.

149. Kashyap M, Seneviratne M, Banda JM, Falconer T, Ryu B, Yoo S, et al. Development and validation of phenotype classifiers across multiple sites in the observational health data sciences and informatics network. Journal of the American Medical Informatics

Association. 2020. pp. 877–883. doi:10.1093/jamia/ocaa032

150. Topaz M, Murga L, Gaddis KM, McDonald MV, Bar-Bachar O, Goldberg Y, et al. Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. J Biomed Inform. 2019;90: 103103.

151. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17: 124–130.

152. McGraw D, Rosati K, Evans B. A policy framework for public health uses of electronic health data. Pharmacoepidemiol Drug Saf. 2012;21 Suppl 1: 18–22.

153. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014;21: 578–582.

154. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216: 574–578.

155. Pfaff ER, Champion J, Bradford RL, Clark M, Xu H, Fecho K, et al. Fast Healthcare Interoperability Resources (FHIR) as a Meta Model to Integrate Common Data Models: Development of a Tool and Quantitative Validation Study. JMIR Medical Informatics. 2019. p. e15199. doi:10.2196/15199

156. Ma X, Imai T, Shinohara E, Kasai S, Kato K, Kagawa R, et al. EHR2CCAS: A framework for mapping EHR to disease knowledge presenting causal chain of disorders – chronic kidney disease example. J Biomed Inform. 03/2021;115: 103692.

157. Liang L, Hou J, Uno H, Cho K, Ma Y, Cai T. Semi-supervised Approach to Event Time Annotation Using Longitudinal Electronic Health Records. arXiv [stat.ME]. 2021. Available: http://arxiv.org/abs/2110.09612

158. Ahuja Y, Wen J, Hong C, Xia Z, Huang S, ScD TC. SAMGEP: A novel method for prediction of phenotype event times using the electronic health record. Research Square. 2021. doi:10.21203/rs.3.rs-1119858/v1

159. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366: 447–453.

160. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27: 2176–2182.

161. Thompson HM, Sharma B, Bhalla S, Boley R, McCluskey C, Dligach D, et al. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. Journal of

the American Medical Informatics Association. 2021. pp. 2393–2403. doi:10.1093/jamia/ocab148

162. Swerdel JN, Hripcsak G, Ryan PB. PheValuator: Development and evaluation of a phenotype algorithm evaluator. J Biomed Inform. 2019;97: 103258.

163. Weaver J, Potvien A, Swerdel J, Voss EA, Hester L, Shoaibi A, et al. Best practices for creating the standardized content of an entry in the OHDSI Phenotype Library. 5th OHDSI Annual Symposium. 2019. Available: https://www.ohdsi.org/wp-content/uploads/2019/09/james-weaver_a_book_in_the_phenotype_library_2019symposium.pdf

164. Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res. 2021;23: e22219.

165. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. BMJ. 2010;341: c4226.

166. Sinnott JA, Dai W, Liao KP, Shaw SY, Ananthakrishnan AN, Gainer VS, et al. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. Hum Genet. 2014;133: 1369–1382.

167. Hubbard RA, Tong J, Duan R, Chen Y. Reducing Bias Due to Outcome Misclassification for Epidemiologic Studies Using EHR-derived Probabilistic Phenotypes. Epidemiology. 07/2020;31: 542–550.

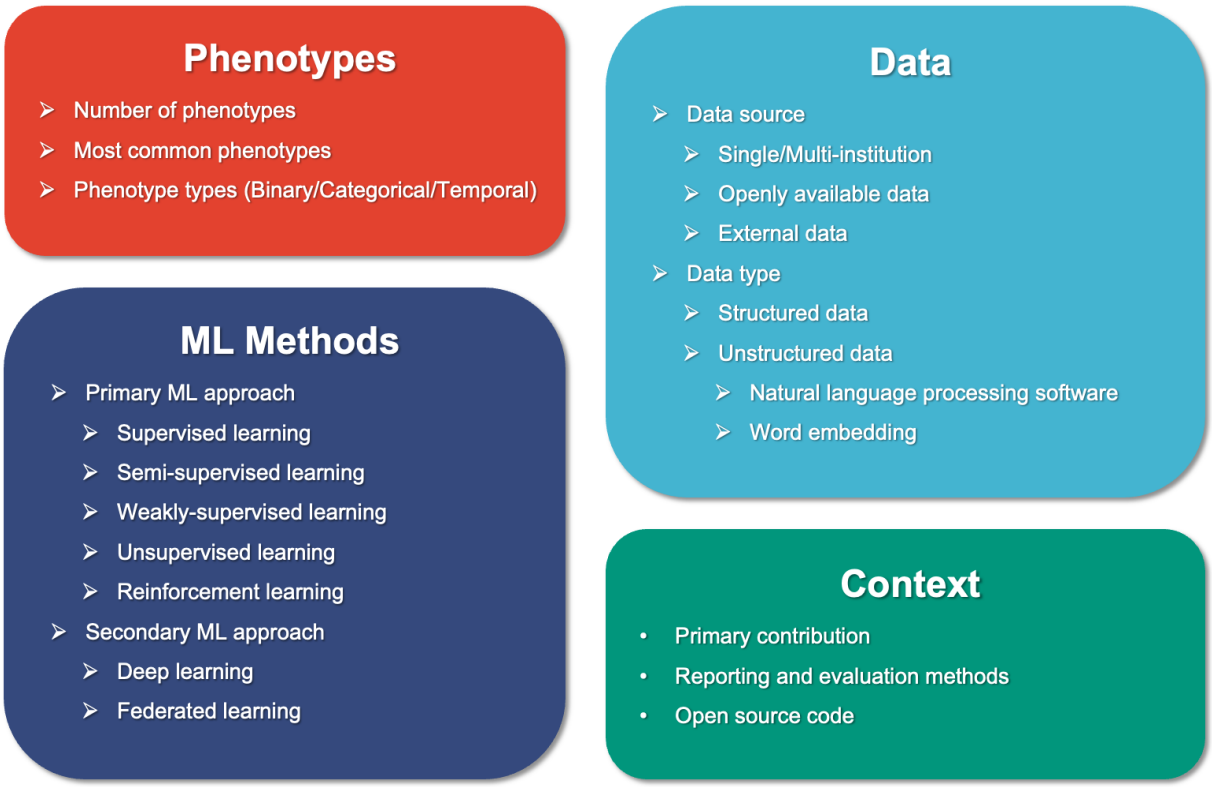

Figure 1. Overview of the four areas considered in our methodical review of the literature.

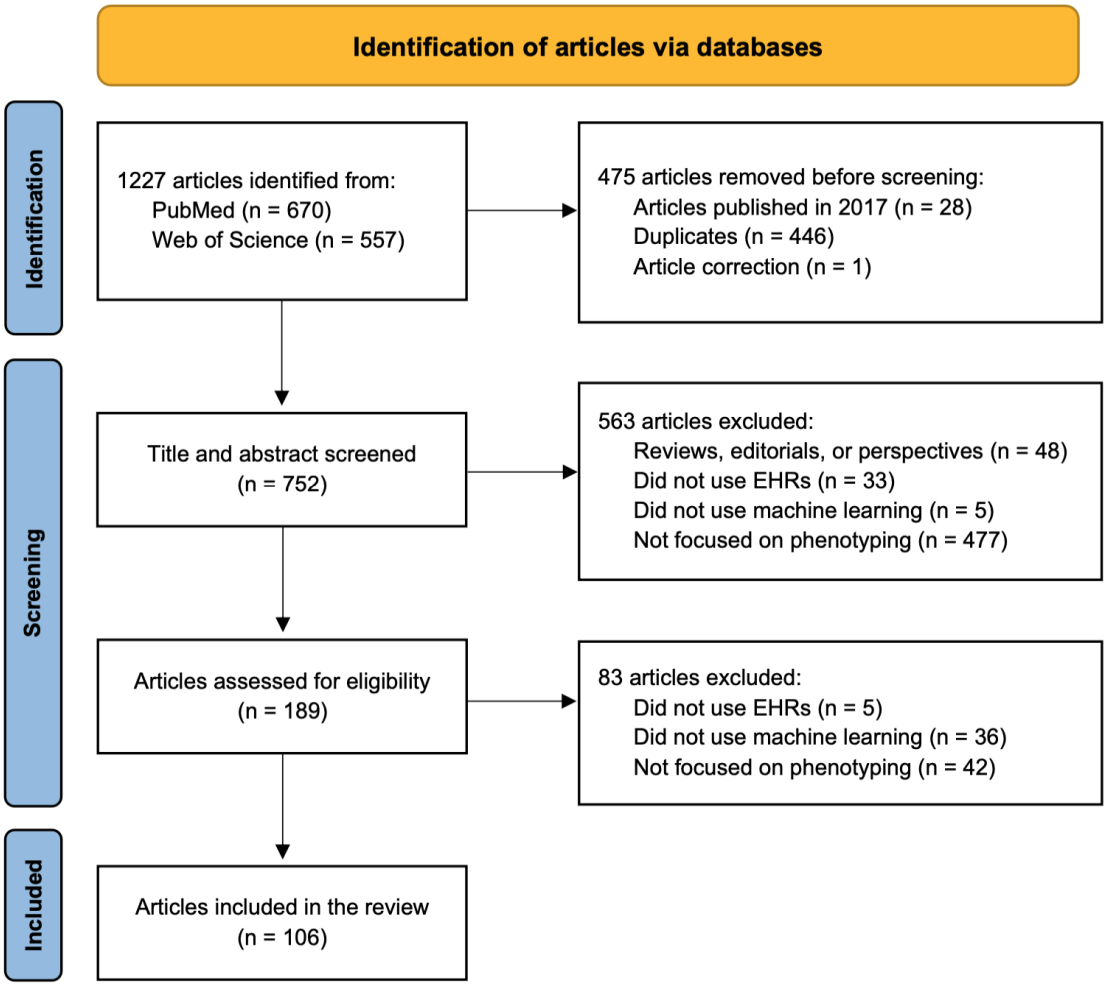

Figure 2. PRISMA diagram for article selection. Only one exclusion reason was chosen for each record during the screening process, although the reasons are not mutually exclusive.

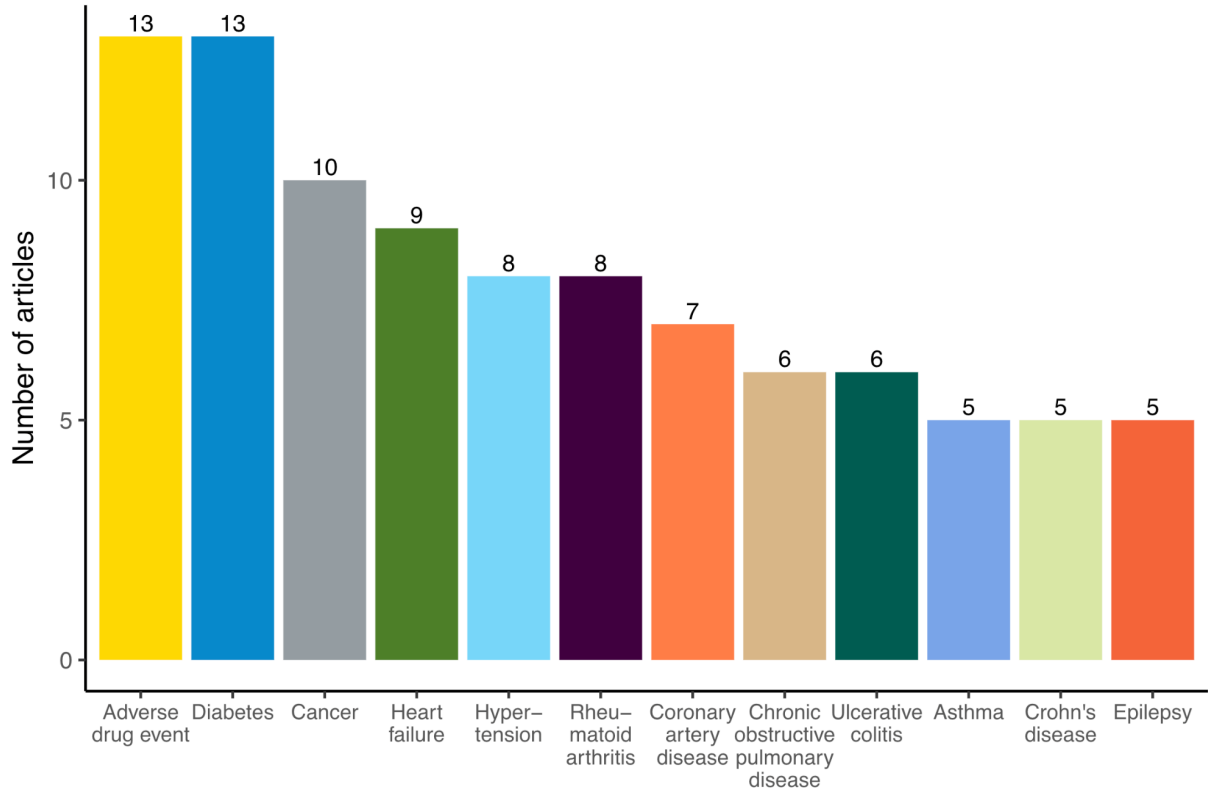

Figure 3. The most common phenotypes within the selected articles.

Note: Phenotypes are presented if they are focused by more than 4 papers. Congestive heart failure (n = 6) and heart failure (n = 3) are counted together, similarly for hypertension (n = 6), resistant hypertension (n = 1), and pulmonary arterial hypertension (n = 1). Cancer includes pan-cancer (n = 1), breast cancer (n = 4), pancreatic cancer (n = 1), ovarian cancer (n = 1), esophageal cancer (n = 1), and metastatic prostate cancer (n = 2).

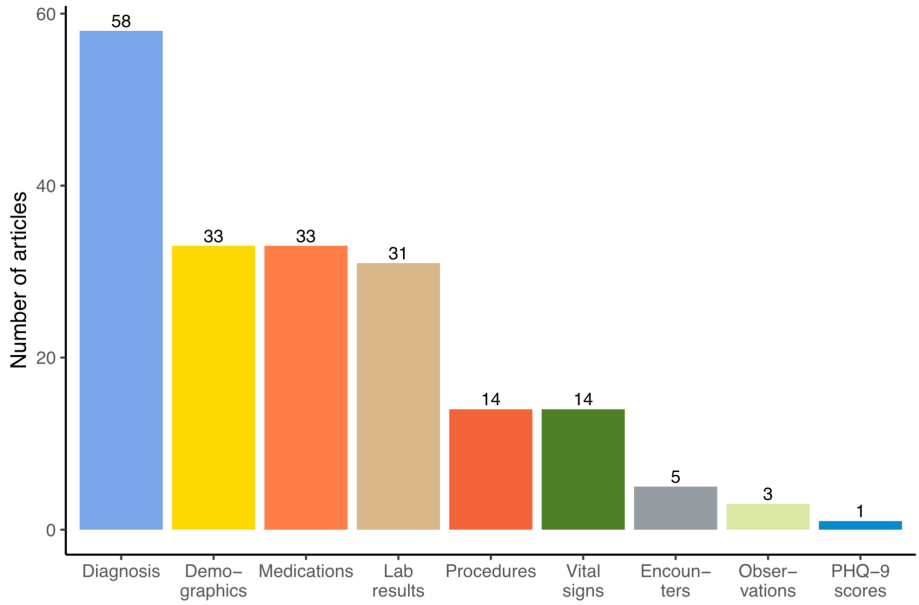

Figure 4a. Types of structured data used in the selected articles.

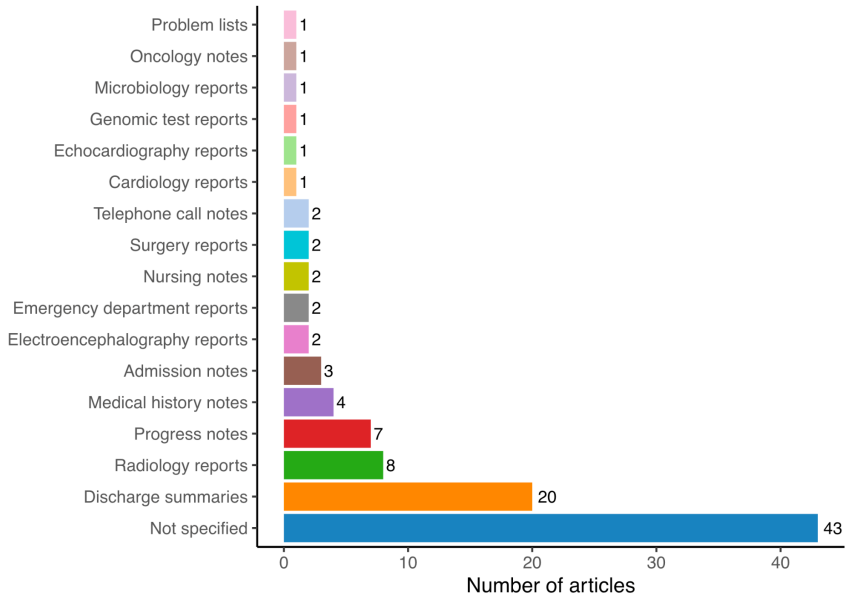

Figure 4b. Types of unstructured data used in the selected articles. Medical history notes include social history, cardiac surgery history, and physical history.

Note: Articles often use more than one type of structured and/or unstructured data.

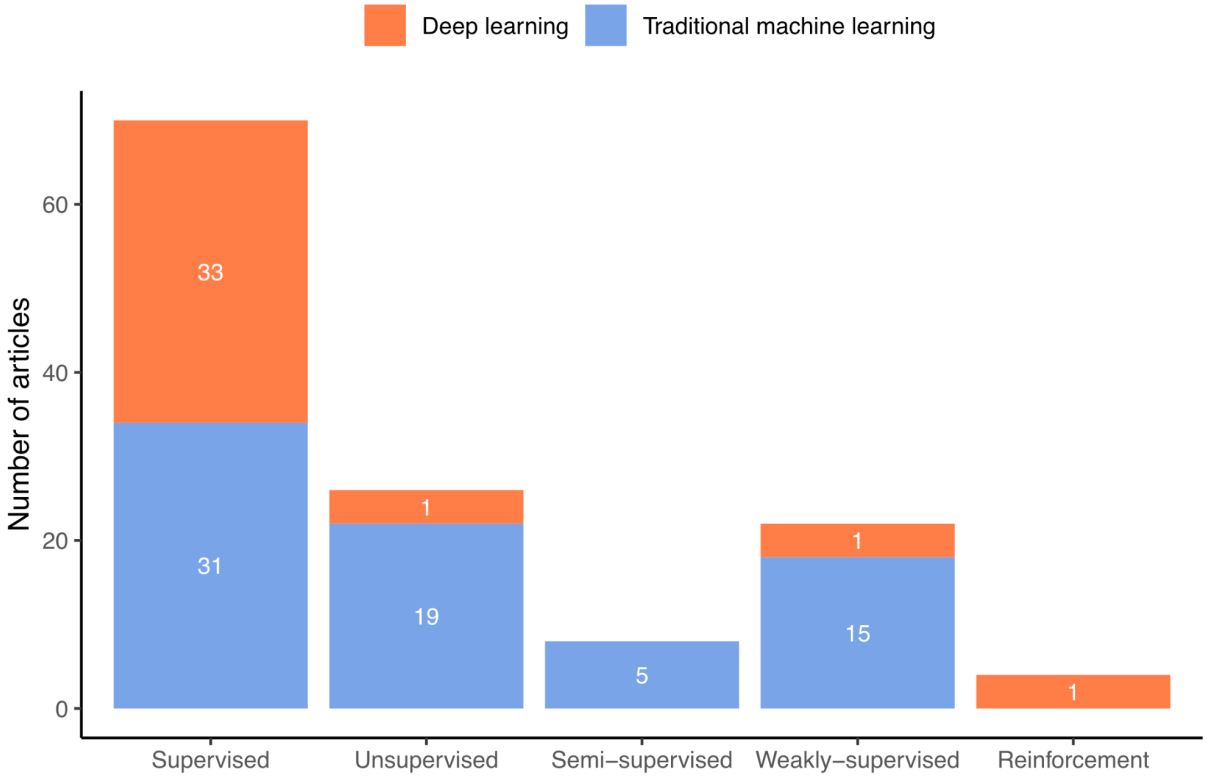

Figure 5. Distribution of machine learning categories among the selected articles and the number of articles using deep learning in each category.

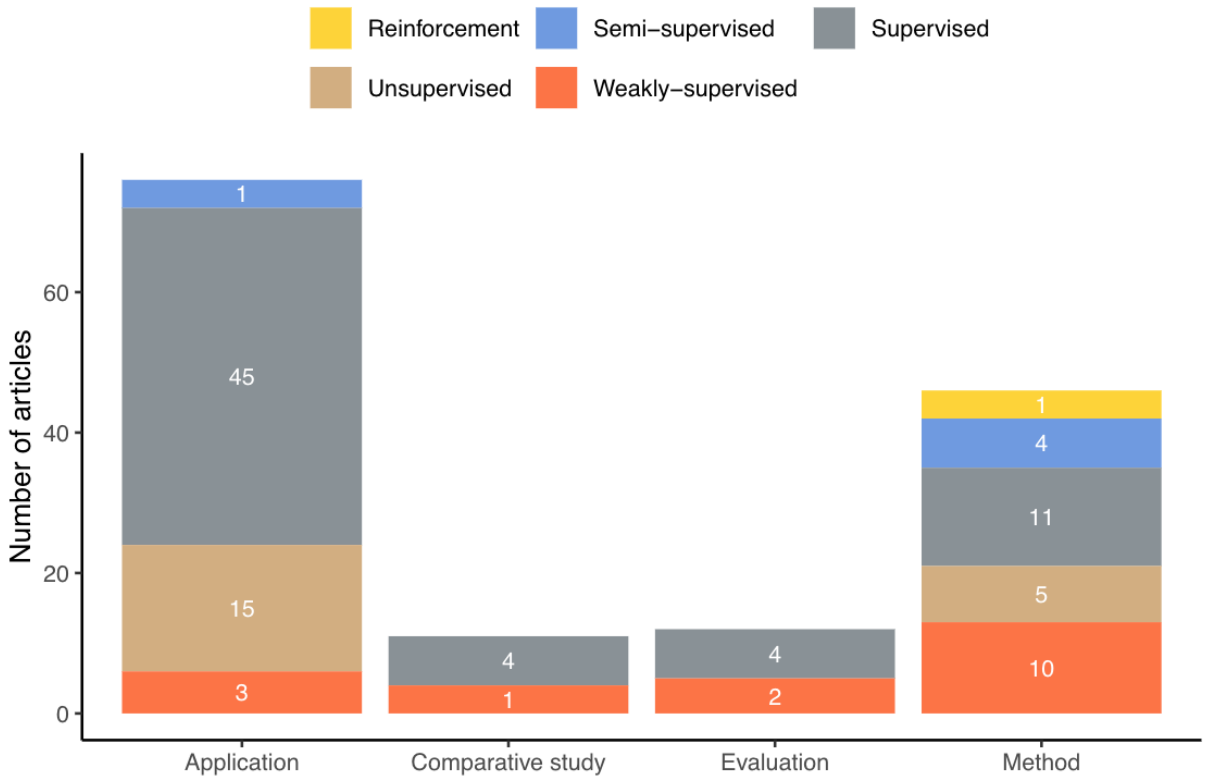

Figure 6. Number of articles and primary contributions.

Table 1. Common prediction performance metrics reported in the selected articles.

Validation metrics	**Number of articles**
Precision (Positive predictive value)	68
Recall (Sensitivity)	66
F-score	49
Area under the receiving operating characteristic curve (AUROC)	48
Specificity	23
Accuracy	18
Negative predictive value	17
Area under the precision-recall curve (AUPRC)	9
Calibration plots	5
Log loss	3
Matthews correlation coefficient	3
Normalized discounted cumulative gain	2
Brier score	2

Note: 82 articles considered multiple evaluation metrics. Metrics are presented if they are used in more than one article. Articles focusing on unsupervised learning use different evaluation methods and are not included in the table.