A review of automatic phenotyping approaches using electronic health records

Alzoubi, Hadeel; Alzubi, Raid; Ramzan, Naeem; West, Daune; Al-Hadhrami, Tawfik; Alazab, Mamoun

Published in:
Electronics

DOI:
10.3390/electronics8111235

Published: 29/10/2019

Document Version
Publisher's PDF, also known as Version of record

electronics

MDPI

Article

A Review of Automatic Phenotyping Approaches using Electronic Health Records

Hadeel Alzoubi [1,*], Raid Alzubi [2], Naeem Ramzan [3], Daune West [3], Tawfik Al-Hadhrami [4,*] and Mamoun Alazab [5]

1 School of Computer and Information Technology, Jordan University of Science and Technology, Irbid 22110, Jordan
2 Department of Computer Science, Faculty of Information Technology, Middle East University, Amman 11831, Jordan; RaidAlzubi@gmail.com
3 School of Engineering and Computing, University of the West of Scotland, Paisley PA1 2BE, UK; naeemramzan@uws.ac.uk (N.R.); daunewest@uws.ac.uk (D.W.)
4 School of Science and Technology, Nottingham Trent University, Nottingham NG11 8NS, UK
5 College of Engineering, IT and Environment, Charles Darwin University, Darwin 0815, Australia; Alazab.m@ieee.org
* Correspondences: hmalzoubi1@just.edu.jo (H.A.); Tawfik.al-hadhrami@ntu.ac.uk (T.A.-H.); Tel.: +44-115-848-4818 (T.A.-H.)

Received: 27 September 2019; Accepted: 22 October 2019; Published: 29 October 2019

Abstract: Electronic Health Records (EHR) are a rich repository of valuable clinical information that exist in primary and secondary care databases. In order to utilize EHRs for medical observational research a range of algorithms for automatically identifying individuals with a specific phenotype have been developed. This review summarizes and offers a critical evaluation of the literature relating to studies conducted into the development of EHR phenotyping systems. This review describes phenotyping systems and techniques based on structured and unstructured EHR data. Articles published on PubMed and Google scholar between 2013 and 2017 have been reviewed, using search terms derived from Medical Subject Headings (MeSH). The popularity of using Natural Language Processing (NLP) techniques in extracting features from narrative text has increased. This increased attention is due to the availability of open source NLP algorithms, combined with accuracy improvement. In this review, Concept extraction is the most popular NLP technique since it has been used by more than 50% of the reviewed papers to extract features from EHR. High-throughput phenotyping systems using unsupervised machine learning techniques have gained more popularity due to their ability to efficiently and automatically extract a phenotype with minimal human effort.

Keywords: electronic health records; phenotyping; natural language processing; machine learning; rule-based

1. Introduction

Electronic health records (EHR) are a rich repository of valuable clinical information that exist in primary and secondary care databases [1]. EHRs are mainly designed to serve patient care. However, with the improvement and development in information technology, increasingly these records are adopted for secondary uses such as quality and safety measurement, clinical decision support, genome-wide association studies (GWAS), and clinical research [1,2]. It has been suggested that such secondary uses of EHR data may be used to improve the quality of health care and reduce medical errors [3]. EHRs comprise of data on a vast range of diseases and health outcomes which enable investigators to conduct high-resolution interventional and observational clinical research.

EHRs contain various types of data that can be categorized into structured and unstructured format. Structured data, such as diagnoses, medication and laboratory results, record patient information using codes and controlled vocabulary. On the other hand, unstructured data, such as radiology reports, pathology reports, discharge summaries and progress notes, are available as narrative text.

In order to leverage EHRs for secondary uses, a range of processes need to be applied on EHR data to transform them into meaningful variables. Phenotyping is one of the most significant processes required to prepare EHR data for research investigations. In this context, phenotyping means identifying a group of patients (cohort) who share a common diagnosis or specific chosen characteristics [4]. Many automatic EHR-based phenotype algorithms were developed retrospectively using rule-based or machine learning techniques [5]. In order to decide if an individual fulfills the phenotype description, an algorithm using features from the EHR needs to be employed [4]. Some phenotype algorithms extract features only from the structured data. However, using these data alone is not always sufficient in developing an accurate phenotyping system. For example, diseases in the field of clinical allergy and immunology lack accurate disease-specific coding which leads to low levels of accuracy [6]. On the other hand, there is an increasing number of phenotyping approaches adopting Natural Language Processing (NLP) algorithms that can extract features from narrative text. These algorithms depend on repeating consecutive steps and finding unobserved relationships within EHR data. Combined features that have been extracted from codified fields and narrative text may increase the sensitivity of the targeted phenotype and improve the overall accuracy of phenotyping systems.

In order to identify the recent articles about developing phenotyping systems from EHR, we followed the NIH Health Care Systems Research Collaboratory [7] and previously published work guidelines [5]. Searches were conducted for articles published between 2013 and 2017 on PubMed and Google scholar, using search terms derived from Medical Subject Headings (MeSH): "phenotype" OR "phenotyping" OR "phenomic" AND "EHR" (OR alternate terms). We included works (including conference proceedings) published in English that met all of the following criteria: 1) Characterizing patients or cohorts with a specific disease or clinical condition, 2) Utilizing EHR data or any clinical reports, 3) Developing or evaluating automatic phenotyping systems. We excluded works that used only manual methods or were not concerned with identifying patient cohorts.

After conducting a comprehensive search, we found that the latest review in the field was in 2014 [5]. In that paper, the authors reviewed studies published between 2010 and 2012 and they focused on the tools and techniques used to build phenotyping systems. Research works since 2013 started focusing more on adopting standardized terminologies in concept extraction and on developing phenotyping systems using unsupervised machine learning and deep learning methods. Another recent closely related study that covers case detection was conducted in 2016 [8]. In that work, Ford et al. present a systematic summary of approaches that used information extraction techniques to improve case-detection. Other works were dedicated to review and evaluate available software tools for authoring EHR-driven phenotyping algorithms [9]. However, the aim of this review is to: 1) provide a summary of the studies that focus on developing or evaluating informatics phenotyping systems based on EHRs, 2) cover the most recent and up-to-date published research papers since 2013, 3) examine the benefits of using many sources of EHR data for developing a phenotyping system, and 4) explore the potential opportunities for future research.

EHRs can support and accelerate observational and interventional clinical research by developing a precise and rapid phenotype definition mechanism [10]. Generally, phenotyping systems follow a sequence of steps in order to identify individuals with a specific disease. Different techniques can be applied in each step based on the dataset structure of the given institution. To help in the structuring and evaluation of the relevant literature, we illustrate the typical structure of a phenotyping system model in Figure 1.

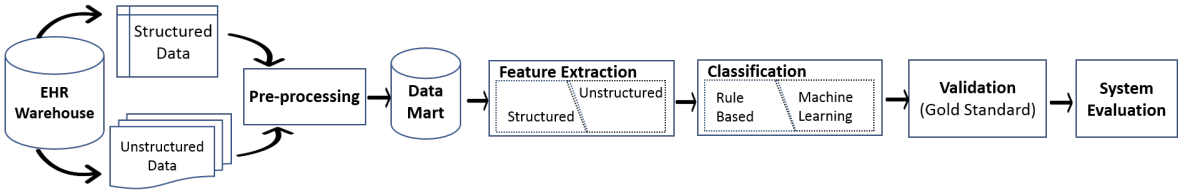

Figure 1. General model of automatic phenotyping algorithms.

The proposed general phenotyping system starts with identifying an EHRs warehouse, as discussed in Section 2. Then, pre-processing is performed in order to identify a "data mart", as discussed in Section 3. Pre-processing is followed by feature extraction where the discriminative features from structured and unstructured fields are extracted, as described in Section 4. The extracted features are injected into a classification system as a third step. Classification methods adopt machine learning techniques or rule-based systems to ascertain individuals with a particular condition (disease), as discussed in Section 5. Finally, the resulting classes are commonly validated against a gold standard and evaluated using different criteria such as sensitivity and positive predictive value, as described in Section 6. Finally, our conclusion is drawn in Section 7.

2. EHR Warehouse

An EHR Warehouse shares data from different health care systems, academic medical centers, and other health-related sources. These warehouses allow research to be embedded within routine care delivery. Since EHRs are mainly designed to serve patient care, the available information does not translate directly to appropriate support for clinical investigators. However, developments in information technology facilitate the use of these records for secondary purposes.

An EHR contains different data types which can be found in a structured or unstructured format. Structured data are easy to analyse and summarize, facilitating storage and enabling sharing of data. However, clinicians find structured data insufficient to describe the patient case precisely [11]. In contrast, narrative text gives physicians the flexibility to write much richer details, observations and concepts about patients' health [11,12]. However, there are many challenges to extracting the buried information from these clinical texts and making it computable and accessible for improving healthcare. Extracting this valuable information from EHR is time-consuming, expensive and labour intensive when it is undertaken manually [13].

Phenotyping systems developed using multiple data sources including both structured and unstructured data have been shown to outperform systems using a single data source [14–18]. Out of the reviewed studies, more than 83% used multiple data sources. We noticed that diagnosis codes, laboratory testing, and medications were still commonly used together. A few studies in the literature have used a single data source which is narrative text. More than 36% have used diagnosis codes along with clinical notes, while 26% have used medication codes in addition to diagnosis codes and clinical notes.

In the context of developing phenotyping algorithms, many impediments are encountered when using EHRs:

1. EHRs contain highly sensitive personal medical data and are subject to strong protection regulations [19,20] which make access to this data difficult.
2. A large number of clinical studies have developed phenotyping algorithms for specific institutions and it is difficult to generalize these algorithms across other institutions.
3. Building an EHR-based phenotyping algorithm requires participation from clinical experts to review subsets within a patient's record to label it as a case (healthy) or a control (affected) and to validate the algorithm output; these efforts take a considerable amount of time [21].
4. EHRs contain noisy data due to missing or incomplete information. Moreover, temporal information may be conflicting or incompatible [22].

5. EHR data are highly relational and multi-modal. The information in EHRs is stored in a complex relational structure, and the information of a single record is spread across multiple tables. Restructuring an EHR database to a more simple structure by performing flattening techniques, such as a "join" operation, may result to loss of information [23].

Some institutions collaborate to build a large EHR warehouses that enable researchers to conduct standardized studies. These studies can support large-scale clinical trials across various health systems while ensuring reproducibility and reliability. For instance, Electronic Medical Records and Genomics (eMERGE) Network [24], Strategic Health IT Advanced Research Project (SHARP) [25], and the National Patient-Centered Clinical Research Network (PCORnet) [26] are examples of collaboration aimed at enhancing research using EHRs across multiple institutions. The Phenotype Knowledge Base (PheKB) [27] is a project started in 2012 by the eMERGE Network, to support sharing and developing phenotyping algorithms for research purposes. Ten studies on the reviewed systems obtained data from the Partners Healthcare Research Patient, five studies from Vanderbilt University and five studies from the Veteran Health Administration. The remaining studies received data from other EHR resources.

3. Pre-Processing

The first step in developing a phenotyping system is to identify a group of instances that have any probability of being connected to a specific disease. The selected subset records are called a "data mart" which form the database of the phenotyping system. Usually a data mart is identified by including patients with at least one diagnostic code for a targeted disease [28,29]. Alternatively, a data mart may include patients who have the disease keyword mentioned in their record [16,30] while other studies use a dataset that has been extracted by a previous phenotyping system as a data mart [14,31–34]. Extra rules may apply on the selected records, for example, a certain age range [18,35] or the period of time over which the patient attended the clinic [36].

Developing a phenotyping system in most cases requires labelled datasets. Labelling patients' records as a 'case ' or 'control' is usually conducted manually by domain experts. While the labelling method is expensive and time-consuming, it is usually conducted over a subset of the data mart. The selected subset of the data mart needs to maintain generalizability and avoid selection biases. Out of the reviewed articles, 94% selected the subset randomly. The labelling method is undertaken again to create a gold standard which is used in the validation stage. Further discussion on this can be found in Section 6.

4. Feature Extraction

Feature extraction is one of the main steps in analysing clinical data and a primary key to success in any rule-based or machine learning phenotyping approach. Feature extraction can be defined as identifying a vector set that represents efficiently the content of information for a cohort while reducing the dimensionality space. Phenotyping systems extract discriminating features to distinguish individuals with a specific condition (disease). Feature extraction can be simple and straightforward, such as when using code frequency as a feature [37]. Additional features can be extracted by applying some rules on the structured data fields, such as abnormal laboratory lipid levels or the number of visits [38]. However, in case the previous techniques fail to increase the accuracy of the prediction systems, NLP is usually applied to extract more features from narrative text. A summary of feature extraction methods is presented at the end of this section.

4.1. Structured Feature Extraction

Many EHR structured fields are stored in coded format. These fields vary across different institutions' datasets. 'Diagnoses', 'medication', and 'procedure' fields are the most common types of data represented in code. To represent diagnoses, the International Classification of Diseases Ninth

Revision (ICD-9) is usually used. In representing procedures, the Current Procedural Terminology (CPT) is mainly used while RxNorm is utilized to code medication. Feature extraction using coded data is available and easy to access and is usually conducted by selecting the code that most likely identifies a cohort with a particular disease. However, since coding in EHRs is usually for billing purposes, it is not always sufficient to support clinical research [39].

The accuracy of code extraction within an EHR depends on whether the code reflects the physician' opinion and whether the physician was able to make a diagnosis correctly [40]. In both cases, often the code fields cannot be fully trusted [41]. Recently, extracting features from coded fields alone is rarely used, and some additional rules on other structured fields are needed. Additional features can be extracted by setting a logical constraint on the structured field data to derive informative features describing phenotype status. These features may take different datatypes such as real numbers, true/false or enumerated categories. For example, the number of clinical visits [38], the number of prescriptions [28] and frequency of ICD-9 code [37] can be counted and stored as real numbers. The existence of some values, such as medication code, prescription code and hospital admission, can be stored as Boolean values (0, 1) [33]. The enumerated categories data field can be divided into different subcategories such as (low, median, and high) for laboratory results, (normal, and abnormal) for blood pressure, or classify patients into certain groups based on age [42]. However, in some cases, using structured data alone results in poor predictability in capturing disease cases, making it necessary to look for more features in unstructured fields.

4.2. Unstructured Feature Extraction

While structured data are available and easy to access, there are limitations in using these data to describe patients' phenotypes. Not all diagnoses and symptoms are coded and, therefore, they may offer a weak clue to the patient's status. Unstructured, or free text EHRs, contain valuable information such as patient history and clinical opinion of this history. An analysis of these clinical notes could offer better information and description of the patient's status; however, the manual review of these notes for large-scale projects is costly and highly time-consuming [43]. To unlock and extract the hidden information from the narrative text efficiently and accurately, NLP techniques are usually applied. NLP is a field of computer-based methods allowing computers to understand and process human (natural) language [44]. The ability to extract critical components from these data and convert them into a structured format could provide great value for clinical research [45]. Narrative text has been utilized to supplement structured data in recognizing particular phenotypes of patients [8,46].

Recent developments in NLP techniques have shown an increasing promise in recognizing and extracting meaningful pieces of information from clinical narrative text [47]. NLP techniques automate the processes required to access the large amount of embedded information in EHRs and consolidate the processes into a coherent structure [48]. From an Information Extraction (IE) perspective, it is essential to pre-process the clinical text by some NLP techniques, such as tokenization, sentence detection, word sense disambiguation, part of speech tagging (POS) and parsing. Higher-level NLP techniques, such as negation, temporality, and entity relations, are crucial for the precise interpretation of the extracted information [13]. Adoption of the previously mentioned techniques varies between reviewed systems and some of these components can be or are omitted. For the task of EHR phenotyping, NLP is commonly applied in three different manners: bag of words, keywords search and concept extraction. Further details on these techniques are given in the following sections.

4.2.1. Bag of Words (BoW)

The Bag of Words model is a simplifying representation used in NLP and information retrieval. In this model the document is represented as a collection of its words, disregarding grammar and words order. In phenotyping systems, the frequency (count) of each word in a document is utilized in feature extraction for classification purposes. This model has been used in a number of studies e.g. [42,49–51] to extract features from EHR documents.

In the BoW model the free text is usually converted to lower case and tokenized into individual words (unigrams). These individual words are treated as features for use within a classification model. Afzal et al. [30] used BoW to represent features when classifying asthma patients. They applied the assertion filter initially, where any words after negations (e.g., 'not', 'no') or speculation (e.g., 'might', 'probable') keywords were discarded. Then, all remaining individual words in the patient's record were treated as features. The same model has been adopted in Afzal et al. [50], with an extra stop words removal step. In Wright et al. [49], the frequency of words are treated as feature vectors for use within a classification model. They built a dictionary of frequently occurring words and the number of times each unique word was used was counted. Any words that occurred more than a threshold number of times was included as a feature in their model. An equal weighting was given for all the words that exceed the threshold. In order to classify patients' records, feature vectors were constructed by using a value of 1 if a lexicon word was present, or 0 if it was not.

A simple improvement on using unigrams was presented in [42] by splitting text to bigrams. This study ignored very high frequency unigrams and bigrams to decrease the number of features. A comparison between unigram, bigram and drug name techniques have been conducted in [52] with the best accuracy being achieved by a combination of them all. However, the BoW model produces a very high number of features which increases classification complexity and decreases the predictive accuracy of the classifier, and which makes performing feature selection, after applying the BoW model, mandatory.

Out of the reviewed papers, it is clearly stated that the BoW model is easy to understand and implement. Moreover, the proposed methods offer high flexibility for customization on a specific dataset. However, the BoW technique produces a huge vector space that severely impacts the sparsity of the document representations [52]. Additionally, the BoW technique ignores the order of words, which highly affects the context since meaning and context can improve the sensitivity of the extraction model [50].

4.2.2. Keywords Search

Some studies identified specific keywords for a given disease to detect phenotyping status [5,29]. Such keyword identification was usually conducted through clinical knowledge or medical domain experts. It is essential to conduct a wide enough search to generate a set of keywords that represent the phenotype of interest. However, using an insufficient keywords set in a searching process could lead to some relevant records being ignored. Generating an inclusive set of keywords that enable a search process to detect all relevant documents, but specific enough to avoid including irrelevant documents, is a challenging process. The keyword search method is especially effective when the word of interest is rare and unique and is not frequently used in a negation form [53].

A lexicon is a group of user-defined keywords assembled in a dictionary. The lexicon can be built based on manual review by relevant field experts or by inspecting domain knowledge. However, developing task-specific dictionaries created by relevant field experts is very time-consuming. Building a lexicon manually by an expert has been widely described in the literature [6,54]. Castro et al. [55] validated the use of EHRs for diagnosing bipolar disorder and identifying control individuals. They used the Health Information Text Extraction (HITEx) tool [56] to extract related concepts from narrative notes. The concepts were identified by clinicians as either consistent or inconsistent with a diagnosis of bipolar disorder. Bellows et al. [29] identified patients with binge eating disorder (BED), using a list of keywords created by clinicians and health information management professionals. ConText [57] was used to detect if BED terms are negated or given to someone other than the patient. Ludvigsson et al. [58] developed an EMR-based algorithm to detect patients at high risk of Celiac Disease (CD) and in need of CD screening. Their algorithm used keyword search methods to extract exactly matched concepts from clinical notes according to the clinical experience of medical practitioners. The developed system showed that a keyword search based method has higher sensitivity than an algorithm based on ICD-9 codes to ascertain high-risk CD patients.

Abhyankar et al. [16] adopted a similar strategy to the one described above, with the additional step of expanding the keyword list by using synonyms obtained from exploring domain knowledge. Gundlapalli et al. [59] obtained the keywords from catheter-associated urinary tract infection (CAUTI) related published work, and a list of terms relevant to CAUTI from Centers for Disease Control and Prevention. The resulting list has been expanded using words from domain experts. In the case of Hanauer et al. [60], the authors extract the keywords from the list of terms that come with the ICD-9 code.

The keyword list is efficient when it contains unique and rare words [53]. However, generating the keyword list is very costly, time consuming and tedious task when it is created by field experts. Moreover, misspelled words and ungrammatical sentences in the medical text can significantly decrease the performance of extraction model [5]. In order to circumvent these limitations, automatic keyword extraction become demanding particularly with appearance of huge medical data [61]

4.2.3. Concept Extraction

Information extraction is a common task in NLP and it is often used to extract associated concepts and organize data from clinical documents. Concept extraction is a significant primary step in extracting meaningful terms and mapping them to standardized coding systems for clinical terminologies such as Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [62], RxNorm, and Unified Medical Language System (UMLS) [63]. Concept extraction automatically identifies and assigns values to terms in the free text which represent concepts such as (disease, symptoms, procedures, or medications). Concept extraction is usually followed by determining negation and uncertainty, understanding temporal relations, and deducing family or social history information. A summary of all studies that used this method to extract unstructured features is presented at the end of this section in Table 1 and further details are discussed below. A large number of tools and frameworks are currently available for clinical information extraction purposes, such as clinical Text Analysis Knowledge Extraction System (cTAKES) [64], MetaMap [65], Medical Language Extraction and Encoding (MedLEE) [66] and KonwledgeMap Concept Identifier (KMCI) [67]. Most frameworks employ a cumulative process pipeline, by moving from basic tasks, like tokenizing words or paragraphs, into higher-level tasks that may depend on the outputs of upstream processes. The cTAKES system, which was developed by the Mayo Clinic, is one of the most popular NLP systems in the literature. This system parses the clinical free text in order to identify the types of relevant clinical concepts in addition to qualifying elements such as (negated or non-negated, current or history). The concept code is provided by mapping each class to a specific terminology domain, which is responsible for handling language variations.

Most previous studies use the cTAKES software to extract the informative concepts from the narrative text. An extended module of the cTAKES system, the "Document Time Relation" has been described in [68]. This extended module can capture the temporal relation between medical events and document creation time. The extracted features from this module were added to the concept unique identifier (CUI)-coded from a customized dictionary, temporal signals, drug attributes, and section parsing. These features were then used to identify patients with methotrexate-related liver toxicity among a rheumatoid arthritis cohort. Xia et al. [28] built a customized dictionary to identify multiple sclerosis patients based on clinician experiences. They used the cTAKES system to annotate the clinical notes and extract narrative variables such as symptoms, signs, and medications. They utilized these variables to refine the dictionary. The final list of terms was mapped to SNOMED-CT and RxNorm. That work determined the features by the sum of positive and negative mentions of narrative variables per patient. Ananthakrishnan et al. [15] used a combination of codified data and clinical notes to identify patients with inflammatory bowel disease (IBD). The codified data were: a) the total number of ICD-9 codes for Crohn' disease and ulcerative colitis, and b) the related medication code of the given diseases. Furthermore, the cTAKES tool was used to extract IBD-related concepts from narrative text notes. The extracted concepts were defined by experts and assigned with SNOMED-CT terminologies.

They deduced that incorporating narrative concepts from EHRs in the developed system allows the identification of more patients than using codified data alone.

Table 1. Summary of studies using concept extraction method.

Disease	Software	Terminology	Extracted Features	Ref
Autism Spectrum Disorder	cTAKES	UMLS (SNOMED-CT and RxNORM) and manual project-specific code	A vector of concept unique identifiers (CUIs)	[69]
Hypertensive	KMCI	UMLS (SNOMED-CT)	Count the appearance of the concept in the note	[70]
Inflammatory Bowel	cTAKES	UMLS	Customized list of narrative concepts	[71]
Rheumatoid Arthritis	NILE	UMLS	Count the appearance of the concept in the note	[37]
Influenza	Topaz, MedLEE	UMLS	Assign a single value of 'present', 'absent', or 'missing' (not mentioned) to each influenza-related findings	[72]
Rheumatoid Arthritis	cTAKES	UMLS (SNOMED-CT and RxNORM)	A vector of frequency of concept unique identifiers (CUIs)	[68]
Lymphoma		UMLS	Relations among a flexible number of medical concepts sentence subgraph features	[73]
Crohn's Disease and Ulcerative Colitis	cTAKES	SNOMED-CT and RxNORM	The number of times the terms were mentioned in the narrative notes and the number of times the terms were mentioned in the clinical text for each subject	[15]
Venous Thromboembolism	KMCI	UMLS	Sentence contains concepts	[74]
Multiple Sclerosis	cTAKES	SNOMED-CT and RxNORM	The sum of positive and negative mentioned concepts per patient.	[28]
Rheumatoid Arthritis	cTAKES	UMLS	A vector of concept unique identifiers	[51]
Blunt facial trauma victims	MedLEE	UMLS	Certainty and temporal status modifiers	[75]
Peripheral Arterial Disease	MedTagger	MedLex	Count the appearance of the concept in the note	[31]
Crohn's Disease, Multiple Sclerosis, Rheumatoid Arthritis and Ulcerative Colitis	cTAKES, HITex	UMLS (SNOMED-CT and RxNORM)	Number of mentions of the concept	[76]
Influenza	Topaz	UMLS	A vector of concept unique identifiers	[77]

Other NLP tools for concept extraction have been used in different studies. For example, Hinz et al. [74] designed a NLP algorithm to capture acute and historical cases of venous thromboembolic disease (VTD) in de-identified EHRs. Initially, they identified the patients using the ICD-9 code while the problem list was used to identify the patients who do not have the ICD-9 code. Then, they employed NegEx [78] to detect classical negation of clinical notes. After that, the KMCI system was utilized to derive a list of concepts from UMLS knowledge resources. Finally, a second round of negation detection was undertaken in order to increase the system overall accuracy. The authors concluded that the use of ICD-9 codes as a precursor for VTD identification is more effective than using NLP techniques alone. Another study was performed on prior diagnostic imaging for blunt facial trauma victims [75]. That study aimed to classify imaging reports of computed tomography from an emergency department. The MedLEE tool was used to extract medical terms and modifiers for certainty and temporal status from reports. This system showed some promising results in automating the outcome classification from free clinical data. Ye et al. [72] compared the Topaz [79] and MedLEE tools for identifying influenza-related findings, and they found that Topaz performed better than MedLEE in terms of accuracy. Yu et al. [37] collected comprehensive medical concepts and grouped drugs concepts based on their relationship in the UMLS. They used the NILE tool [80] to parse clinical notes and identify occurrences of the UMLS concepts. After that, heuristic rules were used to

remove uninformative or highly nonspecific concepts. Finally, the number of positive mentions for each concept, from all notes of each patient, were counted and used in model building. Afzal et al. [31] ascertained peripheral arterial disease (PAD) status using NLP and rule-based methods. The authors utilized the MedTagger tool [81] to process clinical text and annotate clinical concepts. They detected PAD named entities according to a PAD-specific dictionary. The developed dictionary was built by medical experts and then expanded to synonyms by MedLex [82] to discover PAD related concepts. This approach was shown to attain high sensitivity and specificity in identifying PAD cases.

Luo et al. [73] developed an automated interpretable lymphoma classification system, based on unsupervised extraction of relations among medical concepts. This work used narrative clinical sentences in pathology reports to create a sentence subgraph mining model for capturing relations between medical concepts and then generate features. Their phenotyping system used the UMLS metathesaurus to map token subsequences to concepts and, in turn, sentence graph nodes, aiming to ensure meaningful interpretations over the sentence graphs.

By assigning numbers to the clinical symptoms and conditions, the concept extraction technique provides standardization. This technique eliminates the confusion that may result from the use of regional terms. Moreover, the numerical representation of the terms facilitates the exchange of clinical information among different institutions. However, providing comprehensive terminology that includes all clinical terms still a challenging process [83].

4.2.4. Feature Selection

The previously described methods for extracting features from unstructured fields could produce a vast amount of terms which generate a huge feature space. A small subset of features has a strong correlation among each other and with the class label. Therefore, various techniques for feature selection have been successfully applied to reduce the feature space [51,84–88]. It is worth mentioning that chi-square [89] is the most popular feature selection method in the reviewed literature. Extracting features from both structured and unstructured data improves phenotype identification. Moreover, using NLP methods to extract features from narrative text has demonstrated success in the reviewed studies, with concept extraction being the most commonly NLP technique adopted, that has improved standardization in the field due to using standardized terminologies. However, the improvement in the NLP techniques such as co-reference resolution and assertion classification has not been invested adequately in the field of phenotyping systems. A summary of feature extraction methods is presented at the end of this section in Table 2.

Table 2. Categorise of feature extraction techniques.

Feature Extraction Method	Papers
Structured only	[18,35,38,90–96]
BoW only	[30,49]
Keyword search only	[29,54,97]
Concept extraction only	[31,59,69,72,73,75,98]
Structured + BoW	[42,50,52,99]
Structured + Keyword search	[6,16,17,32,55,58,60]
Structured + Concept extraction	[15,28,36,37,48,68,70,71,74,100]

5. Classification

Classification is the process of identifying a group of patients who share a common diagnosis or specific chosen characteristics. This process is usually conducted via a rule-based or machine learning approach. In a rule-based system, a set of rules needs to be developed on the extracted features as a condition that should exist in order to identify patients' records. These rules are either set by a

group of experts in the given field or derived from an organization. By comparison, machine learning techniques aim to develop computer programs that are able to learn by themselves, detect patterns in data, and alter program actions according to new data. Hybrid systems that merge both approaches have been used in some studies [50,93].

5.1. Rule-Based

In the phenotyping context, a rule-based system is a set of logical constraints and rules that can be applied to EHR data to derive the phenotype status of the patients. Rule-based approaches perform well when a task comprises a specific subdomain or a limited number of named entities [101]. Rule-based phenotyping algorithms range from matching a simple pattern to more complicated symbolic approaches. Rule-based systems may comprise multiple logical steps and merge diverse operations such as Boolean (AND, OR and NOT), Comparative (threshold a variable) and Aggregative functions (COUNT, FIRST) [99,102,103]. This technique has the advantage of being easy to construct, accurate when using small datasets, and reliable since it uses human-interpretable strategy. Moreover, it needs only a small number of charts/patients records to be reviewed for training/validating the process. However, developing this technique is time and effort consuming due to the requirement for clinical and informatics knowledge. In this review, we have focused on the way that the rules were generated to build rule-based phenotyping systems. A summary of systems that applied rule-based classification methods is presented in Table 3.

Table 3. List of studies using Rule-based systems as a classification technique.

Criteria	Year	Phenotype	Se	Sp	F1	PPV	NPV	AUC
Expert	2017 [12]	Systemic Lupus Erythematosus	40	-	0.56	91	-	-
	2016 [6]	Aspirin-exacerbated Respiratory Disease	-	-	-	81	-	-
	2016 [71]	Non-Response to Treatment	-	-	-	-	-	0.91
	2016 [32]	Diabetic Retinopathy	-	-	-	75	100	-
	2016 [31]	Peripheral Arterial Disease	96	98	-	92	99	-
	2016 [59]	Indwelling Urinary Catheter	72	-	-	98	-	-
	2016 [98]	Carotid Stenosis	88	84	-	70	95	-
	2015 [104]	Pancreatic Cyst	99	98	-	-	-	-
	2014 [105]	Prostate Biopsy	99	99	-	99	99	-
	2014 [36]	Incident Antidepressant Medication	-	-	-	90	98	-
	2014 [47]	Heart Failure	93	-	0.93	92	-	-
	2014 [29]	Binge Eating Disorder	96	-	-	-	-	-
	2014 [16]	Dialysis	100	98	-	78	-	-
	2013 [74]	Venous Thromboembolism	90	-	0.89	84	-	-
	2013 [15]	Crohn's Disease	69	97	-	98	-	0.95
		Ulcerative Colitis	79	97	-	97	-	0.94
	2013 [58]	Celiac	72	82	0.78	-	-	-
Healthcare Guidelines	2017 [92]	Dyslipidemia	94	-	-	-	79	0.97
	2016 [94]	Liver Injury	80	100	-	-	-	0.95
	2016 [106]	CA-MRSA (Case)	94-100	-	-	68-100	-	-
		CA-MRSA (Control)	75-100	-	-	96-100	-	-
	2016 [107]	Familial Hypercholesterolemia	-	-	-	94	97	-

Where Se is sensitivity, Sp is specificity, F1 is F1-score, PPV is Positive Predictive Value (Precision) , NPV is Negative Predictive Value and AUC is Area Under the Curve.

5.1.1. Rules Based on Clinical Judgment

Most developed systems employed clinical referees (e.g., physicians or other expert opinions) to formulate logical rules. This technique starts with developing an algorithm based on rules that are generated from some expert's view of the targeted domain. Rules are then iteratively improved by experts through chart review and validation by EHR data.

Rule-based phenotyping approaches can be simple and straightforward such as categorizing samples based on keyword matching. For example, Hanauer et al. [60] developed an algorithm to detect patients with disorders of sex development. The potential cases were defined as the patients who had at least one of the keywords which had been specified by clinical experts. Otherwise, clinical rules are offered to guide phenotyping algorithm development. For instance, Michalik et al. [38] developed and validated a computable phenotyping algorithm that identified cohorts with sickle cell disease. The developed algorithm utilized administrative claims data (ICD-9 codes, number of visits, and hospital admissions) to detect cases. The algorithm showed that EHR data can be harnessed to identify patients with sickle cell disease accurately. Restrepo et al. [32] developed an algorithm to identify a diabetic retinopathy cohort among type-2 diabetic African Americans. The algorithm incorporated a combination of diagnostic codes and CPT billing codes to identify cases. In addition, medications and text matching were used to ascertain controls.

5.1.2. Rules Based on Healthcare Guidelines

A few studies use guidelines and recommendations from health institutions to derive rules for specific diagnoses. Oake et al. [92] developed a phenotyping algorithm for identifying patients with dyslipidemia. The rules were built using the ICD-9 code for the given disease and abnormal laboratory lipid levels, which were determined according to Canadian guidelines for the diagnosis and management of dyslipidaemias. The authors used this algorithm as a gold standard to assess and validate their algorithm in later work [35]. Based on the Centers for Disease Control and Prevention (CDC) definition, Jackson et al. [106] developed and validated a Community associated-methicillin resistant Staphylococcus aureus (CA-MRSA) phenotyping algorithm. Some studies have used predefined criteria to build their phenotype models. For example, Safarova et al. [107] developed an automatic phenotyping algorithm to detect patients who met the Dutch Lipid Clinic Network (DLCN) criteria. By comparison, internationally agreed consensus of the importance of biochemical criteria was used in [94] and classification criteria for interstitial pneumonia with autoimmune features was used in [108].

5.2. Machine Learning

With access to clinical data, machine learning (ML) and statistical model-based phenotyping algorithms have been adopted more frequently. These techniques allow programs to deduce patterns from a dataset during the learning phase, which in turn, allows it to generalize predictions about a different dataset. During the training phase, multiple iterative processes are used to optimize numerical parameters that describe a given algorithm's underlying framework. Using ML methods in phenotyping algorithms can reduce the effort needed from clinical domain experts, since there is no need to set rules. In this context, ML approaches are used to enable correct predictions for target diagnoses, based on given observed features from corresponding samples. Different phenotyping systems based on learning algorithms are presented and discussed in this section.

ML approaches can be classified into supervised or unsupervised. In supervised learning, the input data are labelled and the system must learn a pattern from these data to predict the desired output [109]. However, this approach depends on annotated resources which are expensive, difficult, and time consuming to generate in the medical field. In unsupervised learning, the data are not labelled and the system tries automatically to recognize a pattern within input data, such as data clustering [109]. With the availability of huge amounts of unannotated data, recently this technique

has gain increased popularity. However, using unsupervised methods could produce various clusters of EHR that may not be related to the targeted phenotypes, since there are no training examples to guide the clustering process [110].

5.2.1. Supervised Learning

Supervised learning aims to construct predictors that distinguish individuals with a specific disease (case) from unaffected ones (control). The main pitfall in supervised methods is the possibility of over-fitting: the model tends to fit the learning samples data perfectly but does not predict well for unseen data. Over-fitting happens when the model learns the random noise in the learning phase instead of learning only from desired features. The risk of over-fitting can be avoided by applying a cross-validation technique which helps to provide a more objective evaluation of the performance in unseen cases. The most popular supervised techniques that were used in the literature are Support Vector Machine (SVM), Decision Tree, and multiple logistic regression methods [111,112]. Table 4 presents the classification algorithms and evaluation metrics for studies that used supervised learning as a classification technique.

Wright et al. [49] used the SVM-based approach for classifying EHR progress notes about diabetes. The developed model showed an ability to generalise as it was able to perform properly over different institutional datasets. Experiments over a range of ML algorithms and features were conducted in [51]. The best performing algorithm was linear kernel SVM with AUC 0.83%. Cohen et al. [52] compared the Naive Bayes classifier and SVM to classify potential candidates for surgical intervention for drug-resistant pediatric epilepsy, with 10-fold cross-validation. The authors concluded that SVM achieved better performance than the Naive Bayes classifier with an F1-score of 0.77 for Naive Bayes and 0.82 for SVM.

Various decision tree algorithms have been used in previous studies. For example, Yadav et al. [75] used CART decision tree modelling as a classification technique on de-identified computed tomography reports, after the dataset was randomly divided into training and testing sets. Zhou et al. [95] used the C5.0 decision tree, which is an improved version of C4.5, to identify patients with rheumatoid arthritis from primary care EHRs. The authors deduced that ML methods can be utilized to create reliable disease phenotypes in EHRs. Wu et al. [97] developed an automated algorithm for asthma ascertainment from EHRs. The C4.5 decision tree algorithm with 10-fold cross-validation was used to classify patients and detect asthma status. The system showed that the ML algorithm is better than the rule-based method in terms of the F1-score which increased from 0.82 to 0.86.

Logistic regression has been used in different studies. For example, Fan et al. [91] constructed and validated an algorithm to identify peripheral arterial disease (PAD) status using administrative data. The training set was used to estimate an integer for billing codes score by applied multiple logistic regression methods, and these codes were tested in the validation set. The developed system demonstrated reasonable accuracy in identifying PAD cases. Castro et al. [55] and Xia et al. [28] have applied LASSO penalized logistic regression models with Bayesian Information Criterion in order to select the informative variables and build a predictive model successfully.

Afzal et al. [30] used four ML approaches (C4.5, SVM, RIPPER, MyC) to build an automatic case identification system for hepatobiliary disease and acute renal failure. This study focused on improving the sensitivity of classifiers, so it used sampling and cost-sensitive learning techniques to deal with the imbalance between positive and negative examples. The authors utilized 5-fold cross-validation to train and test the classifiers and they found no clear differences in their system accuracy after applying various ML techniques. Garg et al. [42] undertook a comparison between different ML techniques. In their comparative study the best method was ensemble classification with the highest average F1-score of 0.98. Pineda et al. [77] compared the diagnostic capabilities of seven ML classifiers for influenza detection against an expert-built influenza Bayesian classifier. They used different techniques to handle the missing data from the clinical reports in order to increase the predictive accuracy. This study

showed that all machine learning classifiers have an area under curve (AUC) ranging from 0.88 to 0.93, and performed significantly better than the expert-built model.

Deep learning is another ML technique that has been used to identify patterns for phenotyping purposes. The main difference between ML and deep learning methods is that deep learning can learn features from big data, rather than applying feature selection techniques. Deep learning methods have the advantage of automatically learning form huge feature space such as EHR. Lasko et al. [113] introduced a deep learning method for phenotype discovery in clinical data. Their approach couples dirty clinical data with a deep learning architecture via longitudinal probability densities inferred using Gaussian process regression, achieving an AUC of 0.97. Lipton et al. [114] applied Recurrent Neural Networks (RNNs) to identify 128 diagnosis, their system outperforming multilayer perceptrons. A comparison of rule-based and convolutional neural networks (CNNs) approaches for patient phenotyping has been conducted in [115]. They concluded that deep learning is an effective approach to build a phenotyping system based on clinical notes. Deep learning has also been applied in [116] using autoencoder.

Table 4. List of studies using supervised learning as a classification technique.

Paper name	Phenotype	ML Method	Se	Sp	F1	PPV	NPV	AUC
2016 [42]	Cardiac Amyloidosis	K-NN, SVM, decision tree, Random Forests, AdaBoost, and Naïve Bayes	-	-	0.98	-	-	-
2016 [70]	Hypertensive	Random Forests	90	-	-	95	-	0.97
2016 [52]	Epilepsy	SVM, Naive Bayes	-	-	0.78	-	-	0.83
2016 [95]	Rheumatoid Arthritis	C5.0 decision tree	93	99	-	90	-	-
2016 [96]	Type 2 Diabetes	Multivariate Logistic Regression and a Random-Forests Probabilistic Model	80	74	-	40	94	0.84
	Inflammatory Bowel Disease (Crohn's Disease)		72	-	-	98	-	-
	Inflammatory Bowel Disease (Ulcerative Colitis)		73	-	-	97	-	-
2015 [76]	Multiple Sclerosis	Logistic Regression	78	-	-	95	-	-
	Rheumatoid Arthritis		63	-	-	94	-	-
2015 [37]	Rheumatoid Arthritis	Penalized Logistic Regression	-	-	0.77	70	-	0.95
	Coronary Artery Disease		-	-	0.82	84	-	0.92
2015 [117]	Gout Flares	SVM	82	92	0.87	77	93	-
2014 [72]	Influenza	Bayesian Network	-	-	-	-	-	0.73
2013 [75]	Blunt Facial Trauma Victims	CART Decision Tree	93	97	0.97	81	99	-
2013 [91]	Peripheral Arterial Disease	Logistic Regression	68	87	-	75	83	0.91
2013 [49]	Diabetes	SVM	92	-	0.93	95	-	0.94
2013 [97]	Asthma	C4.5 Decision Tree	84	96	0.86	88	95	-
2013 [51]	Rheumatoid Arthritis	Logistic Regression , Naive Bayes, Multilayer perceptron , SVM	-	-	-	-	-	0.83
2013 [28]	Multiple Sclerosis	LASSO Penalized Logistic Regression Models with Bayesian Information Criterion	83	95	-	92	89	0.95
2013 [33]	Type 2 Diabetes Mellitus	Association Rule Mining, DT, Logistic Regression, SVM	96	-	0.91	86	-	-
2013 [30]	Hepatobiliary Disease	C4.5, SVM, RIPPER, MyC	95	56	-	-	-	-
	Acute Renal Failure		86	77	-	-	-	-

Where Se is sensitivity, Sp is specificity, F1 is F1-score , PPV is Positive Predictive Value (Precision) , NPV is Negative Predictive Value and AUC is Area Under the Curve.

In spite of the significant usefulness of applying ML in the previously mentioned studies, there is some work that casts doubt on the superiority of ML over rule-based approaches. For example, Lingren et al. [69] developed an automated algorithm for determining an Autism Spectrum Disorder patient cohort from EHRs. They compared the expert-rule-based system and SVM-learning system, concluding that the rule-based algorithm produced a better Positive predictive value (PPV) compared

with ML. Likewise, the rule-based system in [47] outperformed the ML method. We believe that this discrepancy in results comes from the diversity of the EHRs dataset structure and content.

5.2.2. Unsupervised Learning

Unsupervised learning offers methods to cluster EHR records into patient groups regarding phenotypes or subtypes. Unsupervised learning methods can identify patterns that, collectively, form a compact and meaningful representation of the original data, with no need for expert input or labelled examples. However, result validation for phenotypic groups with this approach is challenging, due to the lack of given ground truth on these groups. Unsupervised learning methods can process very large volumes of data and do not require manual labelling, so it is an approach often used when building high-throughput phenotyping systems. Various unsupervised learning methods have been developed and applied to EHRs [118–122]

Tensor factorization is an unsupervised learning technique that has been applied widely to EHRs. Ho et al. [123] developed a non-negative tensor factorization model to derive phenotype candidates without supervision. They investigated the interaction of diagnoses and medications among patients. This study used heart failure as a case study to demonstrate the capability of the model. The system has shown the robustness, stability, and the conciseness in generating high-throughput phenotyping. Ho et al. [124] also developed a non-negative sparse tensor factorization model. The observed tensor was decomposed into two terms, an interaction tensor and a bias tensor. The bias tensor denotes the baseline characteristics that are common amongst the overall population and the interaction tensor defines the phenotypes. The same technique has been employed again in Wang et al. [125], who present pairwise constraints in the formulation to guarantee separate phenotypes. The proposed method handled noisy and missing information in EHR tensors carefully. Another unsupervised technique is Generalized Low Rank Modeling (GLRM) which was used by Schuler et al. [126]. They used two datasets with different characteristics to overcome EHR phenotyping barriers such as missing data, sparsity, and data heterogeneity.

5.2.3. Combined Approaches

Combined systems usually merge both rule-based and machine learning techniques to increase the overall sensitivity of the system. In some cases, the developers intended to utilize ML algorithms, but they found no clear distinct separation between some subclasses. For example, Afzal et al. [50] generated a three-level classification model (two ML classifiers, and one rule-based system). The first classifier classifies definite asthma cases from all other cases, and the second classifier classifies probable and doubtful asthma cases from non-asthma cases. The third level of classification distinguishes between probable and doubtful cases. In this level, rules-based techniques were implemented because the ML failed to distinguish between the remaining cases. Moreover, Nguyen and Patrick [127] built an automated system to classify reportable and non-reportable cancer cases from radiology reports. The developed system combined SVM classifier and rule-based methods to improve the overall performance.

Another combined approach case exists when ML is used to enhance a rule-based classifier. For example, Kagawa et al. [93] employed expert-rules to clearly separate control patients and improve accuracy by using ML to classify complicated cases.

Previous works have shown success in detecting phenotypes from EHR data. However, using supervised ML techniques requires manually labeled cases and controls derived from domain experts review. The required human effort limits the number of patients records that are used in the training phase. As a result, the systems miss the opportunity to learn sufficiently. Moreover, these efforts focus on one or two phenotypes. On the other hand, in order to minimize domain experts effort and potentially increase phenotyping accuracy, building high-throughput phenotypes using unsupervised machine learning techniques is a promising approach, since these techniques can be applied to the whole dataset, decrease human effort and detect various phenotypes at the same time.

6. Validation and Evaluation

Validation usually implies a comparison of definitions against a gold standard. A gold standard is one of the main prerequisites for proper validation of phenotyping systems. It is the most rigorous classification technique for estimating the validity of phenotyping algorithms by correctly detecting individuals with and without the targeted disease [4]. Most studies presented in this review compared the results of developed systems against a gold standard to assess their phenotyping frameworks. Preparation of a gold standard is a resource-intensive operation that needs a careful manual review of clinical data. Developing high-quality gold standard annotation requires a multiple process to label instances with the true state of disease: (i) define a representative context dataset, such as clinical notes, (ii) create standard annotation guidelines and schemes, (iii) annotate patients' records manually by pairs of clinical reviewers, where the annotating process is conducted independently, (iv) require expert clinicians to review any cases where disagreement between the reviewers in the previous stage occurred, and (v) express the reliability of annotations using the Cohen's kappa coefficient agreement which is a statistical method to measure inter-agreement and track the quality of the gold standard [39].

The complication of defining a gold standard technique is related to the quality of EHRs data. While in some mental disorders, a sufficient representation of the patient case comes from an expert clinician, other diseases need a wide range of clinical data sources to determine actual phenotype status.

The performance of automated phenotype algorithms is commonly measured using various information retrieval metrics, such as sensitivity, specificity, positive predictive value, F1-score and Area Under the Curve(AUC), as presented in Table 5.

Table 5. List of evaluation metrics commonly used by phenotyping system.

Performance metrics	Definition	Equation
Sensitivity (Recall) (Se)	The proportion of all actually positive samples that are correctly detected.	$Se = \frac{TP}{TP+FN} \cdot 100$
Specificity (Sp)	The proportion of all actually negative samples that are correctly detected.	$Sp = \frac{TN}{TN+FP} \cdot 100$
Positive Predictive Value (precision) (PPV)	The proportion of positively detected samples that are true positive.	$PPV = \frac{TP}{TP+FP} \cdot 100$
Negative Predictive Value (NPV)	The proportion of negatively detected samples that are true negative.	$NPV = \frac{TN}{TN+FN} \cdot 100$
F1-score (F1)	The weighted harmonic mean of precision and recall.	$F1 = 2 \cdot \frac{Pre \cdot Re}{Pre+Re}$
Area under the ROC (AUROC)	ROC is the graph that represent the trade-off between sensitivity and specificity and area under the curve is equal the probability that a predictor will rank a randomly chosen positive sample higher than a randomly chosen negative one	

Furthermore, for evaluation purposes, most of the studies (78%) select the training and testing subsets randomly. The training subset is used to train the system while the test subset is used to evaluate the final system and report the results. Some studies (20%) utilize cross-validation techniques as an alternative approach to reporting results. K-fold cross validation is the classical approach that randomly partitions the dataset into K subsets. In each case the classifier uses one subset as the testing

subset and the remaining K-1 subsets to train the classifier. The 10-fold and 5-fold cross-validation are the most commonly used methods in phenotyping studies.

7. Conclusions

This review summarizes and offers a critical evaluation of the literature related to studies conducted for the development of EHR phenotyping systems. Using multiple EHR data sources that include both structured and unstructured data has significantly improved phenotypes detection. Fifty-one articles describing phenotyping systems were identified and reviewed. For feature extraction, 12 used structured data, 16 used unstructured data, and 23 used both data types. For classification, 22 used rule-based techniques, 22 used machine learning techniques, while the rest used both techniques to classify individuals. Most of the reviewed studies (83%) used a combination of multiple EHR data sources. NLP techniques have been increasingly used for extracting embedded information from narrative text in order to improve the overall accuracy of the phenotyping systems. Throughout the time period covered by this review, concept extraction was the most popular NLP technique, which has been used by more than 50% of the reviewed papers to extract features from EHR. High-throughput phenotyping systems using unsupervised machine learning techniques have gained more popularity due to their ability to efficiently and automatically extract phenotypes with minimal human efforts. Furthermore, most reviewed models adopted standardization in the reporting of systems performance. Approximately 80% of these models reported sensitivity (recall) and PPV (precision), making the studies more compatible and comparable. The application of NLP in general text has been accelerating over the past several years. However, this growth has not been reflected sufficiently on the medical field, due to some challenges that exist in the clinical notes. For instance, limited access to a shared dataset, difficulties in annotating medical notes and, shortage of annotation standards [128]. Future works should pay more attention to addressing these issues by facilitating access to a common database to improve comparability. Moreover, encourage collaboration and developing NLP standards.

Author Contributions: Conceptualization, H.A., R.A. and T.A.-H.; methodology, H.A., R.A., T.A.-H., M.A., and N.R..; software validation, H.A., R.A. T.A.-H., N.R. and M.A.; formal analysis, H.A., R.A.. and T.A.-H.; investigation H.A., R.A. and T.A.-H.; resources N.R. D.W., and T.A.-H.; data curation, H.A., R.A. T.A.-H., M.A. and N.R.; writing—original draft preparation, H.A., .A.; writing—review and editing, H.A., R.A., T.A.-H., M.A., N.R., and D.W.; visualization, H.A., R.A. and T.A.-H.; supervision, N.R., D.W., .A. and T.A.-H.; project administration, N.R., D.W., M.A. and T.A.-H.; funding acquisition, M.A. and T.A.-H.

Funding: This research was funded by School of Science and Technology, Nottingham Trent University and College of Engineering, IT and Environment, Charles Darwin University.

Acknowledgments: We would like to thank the jointly with the NICS group at School of Science and Technology, Nottingham Trent University, United Kingdom and College of Engineering, IT and Environment, Charles Darwin University, Australia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Häyrinen, K.; Saranto, K.; Nykänen, P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *Int. J. Med. Inform.* **2008**, *77*, 291–304. [CrossRef] [PubMed]
2. Hersh, W.R.; Weiner, M.G.; Embi, P.J.; Logan, J.R.; Payne, P.R.; Bernstam, E.V.; Lehmann, H.P.; Hripcsak, G.; Hartzog, T.H.; Cimino, J.J.; etal. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med. Care* **2013**, *51*, S30. [CrossRef]
3. Botsis, T.; Hartvigsen, G.; Chen, F.; Weng, C. Secondary use of EHR: Data quality issues and informatics opportunities. *Summit Transl. Bioinform.* **2010**, *2010*, 1.
4. Richesson, R.; Smerek, M. Electronic Health Records-Based Phenotyping. *Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials*. Available online: http://sites.duke.edu/rethinkingclinicaltrials/informed-consent-in-pragmatic-clinical-trials/ (accessed on 22 October 2019).

5. Shivade, C.; Raghavan, P.; Fosler-Lussier, E.; Embi, P.J.; Elhadad, N.; Johnson, S.B.; Lai, A.M. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 221–230. [CrossRef]
6. Cahill, K.N.; Johns, C.B.; Cui, J.; Wickner, P.; Bates, D.W.; Laidlaw, T.M.; Beeler, P.E. Automated identification of an aspirin-exacerbated respiratory disease cohort. *J. Allergy Clin. Immunol.* **2017**, *139*, 819–825. [CrossRef]
7. NIH. Suggestions for Identifying Phenotype Definitions Used in Published Research @ONLINE. Available online: https://www.nihcollaboratory.org/Products/Phenotype_lit_search_suggestions_02-18-2014.pdf (accessed on 10 October 2017).
8. Ford, E.; Carroll, J.A.; Smith, H.E.; Scott, D.; Cassell, J.A. Extracting information from the text of electronic medical records to improve case detection: A systematic review. *J. Med. Inform.* **2016**, *23*, 1007–1015. [CrossRef]
9. Xu, J.; Rasmussen, L.V.; Shaw, P.L.; Jiang, G.; Kiefer, R.C.; Mo, H.; Pacheco, J.A.; Speltz, P.; Zhu, Q.; Denny, J.C.; et al. Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational research. *Int. J. Med. Inform.* **2015**, *22*, 1251–1260. [CrossRef]
10. Hripcsak, G.; Albers, D.J. Next-generation phenotyping of electronic health records. *Int. J. Med. Inform.* **2013**, *20*, 117–121. [CrossRef]
11. Ford, E.; Nicholson, A.; Koeling, R.; Tate, A.R.; Carroll, J.; Axelrod, L.; Smith, H.E.; Rait, G.; Davies, K.A.; Petersen, I.; et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: What information is hidden in free text? *BMC Med. Res. Methodol.* **2013**, *13*, 105. [CrossRef] [PubMed]
12. Barnado, A.; Casey, C.; Carroll, R.J.; Wheless, L.; Denny, J.C.; Crofford, L.J. Developing Electronic Health Record Algorithms That Accurately Identify Patients With Systemic Lupus Erythematosus. *Arthritis Care Res.* **2017**, *69*, 687–693. [CrossRef] [PubMed]
13. Meystre, S.M.; Savova, G.K.; Kipper-Schuler, K.C.; Hurdle, J.F. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearb. Med. Inf.* **2008**, *35*, 44.
14. Liao, K.P.; Ananthakrishnan, A.N.; Kumar, V.; Xia, Z.; Cagan, A.; Gainer, V.S.; Goryachev, S.; Chen, P.; Savova, G.K.; Agniel, D.; 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* **2015**, *10*, e0136651. [CrossRef] [PubMed]
15. Ananthakrishnan, A.N.; Cai, T.; Savova, G.; Cheng, S.C.; Chen, P.; Perez, R.G.; Gainer, V.S.; Murphy, S.N.; Szolovits, P.; Xia, Z.; et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: A novel informatics approach. *Inflamm. Bowel Dis.* **2013**, *19*, 1411. [CrossRef] [PubMed]
16. Abhyankar, S.; Demner-Fushman, D.; Callaghan, F.M.; McDonald, C.J. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 801–807. [CrossRef] [PubMed]
17. Wei, W.Q.; Teixeira, P.L.; Mo, H.; Cronin, R.M.; Warner, J.L.; Denny, J.C. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Inform. Assoc.* **2015**, *23*, e20–e27. [CrossRef] [PubMed]
18. Morley, K.I.; Wallace, J.; Denaxas, S.C.; Hunter, R.J.; Patel, R.S.; Perel, P.; Shah, A.D.; Timmis, A.D.; Schilling, R.J.; Hemingway, H. Defining disease phenotypes using national linked electronic health records: A case study of atrial fibrillation. *PLoS ONE* **2014**, *9*, e110900. [CrossRef] [PubMed]
19. Glock, J.; Herold, R.; Pommerening, K. Personal identifiers in medical research networks: Evaluation of the personal identifier generator in the Competence Network Paediatric Oncology and Haematology. *GMS Medizinische Informatik Biometrie und Epidemiologie* **2006**, *2*, 06.
20. Feldman, H.; Reti, S.; Kaldany, E.; Safran, C. Deployment of a highly secure clinical data repository in an insecure international environment. *Stud. Health Technol. Inform.* **2010**, *160*, 869–873.
21. Pathak, J.; Kho, A.N.; Denny, J.C. Electronic health records-driven phenotyping: Challenges, recent advances, and perspectives. *J. Am. Med. Inform. Assoc.* **2013**, *20*, e206–e211. [CrossRef]
22. Peissig, P.L.; Costa, V.S.; Caldwell, M.D.; Rottscheit, C.; Berg, R.L.; Mendonca, E.A.; Page, D. Relational machine learning for electronic health record-driven phenotyping. *J. Biomed. Inform.* **2014**, *52*, 260–270. [CrossRef]

23. Koller, D.; Friedman, N.; Džeroski, S.; Sutton, C.; McCallum, A.; Pfeffer, A.; Abbeel, P.; Wong, M.F.; Heckerman, D.; Meek, C.; et al. *Introduction to Statistical Relational Learning*; MIT Press: Cambridge, UK, 2007.
24. McCarty, C.A.; Chisholm, R.L.; Chute, C.G.; Kullo, I.J.; Jarvik, G.P.; Larson, E.B.; Li, R.; Masys, D.R.; Ritchie, M.D.; Roden, D.M.; et al. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genom.* **2011**, *4*, 13. [CrossRef] [PubMed]
25. Chute, C.G.; Pathak, J.; Savova, G.K.; Bailey, K.R.; Schor, M.I.; Hart, L.A.; Beebe, C.E.; Huff, S.M. The SHARPn project on secondary use of Electronic Medical Record data: Progress, plans, and possibilities. In *AMIA Annual Symposium Proceedings*; American Medical Informatics Association: Bethesda, MD, USA, 2011; pp. 248–256.
26. Collins, F.S.; Hudson, K.L.; Briggs, J.P.; Lauer, M.S. PCORnet: Turning a dream into reality. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 576–577. [CrossRef] [PubMed]
27. Newton, K.M.; Peissig, P.L.; Kho, A.N.; Bielinski, S.J.; Berg, R.L.; Choudhary, V.; Basford, M.; Chute, C.G.; Kullo, I.J.; Li, R.; 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–e154. [CrossRef] [PubMed]
28. Xia, Z.; Secor, E.; Chibnik, L.B.; Bove, R.M.; Cheng, S.; Chitnis, T.; Cagan, A.; Gainer, V.S.; Chen, P.J.; Liao, K.P.; et al. Modeling disease severity in multiple sclerosis using electronic health records. *PLoS ONE* **2013**, *8*, e78927. [CrossRef] [PubMed]
29. Bellows, B.K.; LaFleur, J.; Kamauu, A.W.; Ginter, T.; Forbush, T.B.; Agbor, S.; Supina, D.; Hodgkins, P.; DuVall, S.L. Automated identification of patients with a diagnosis of binge eating disorder from narrative electronic health records. *J. Am. Med. Inform. Assoc.* **2014**, *21*, e163–e168. [CrossRef] [PubMed]
30. Afzal, Z.; Schuemie, M.J.; van Blijderveen, J.C.; Sen, E.F.; Sturkenboom, M.C.; Kors, J.A. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. *BMC Med. Inform. Decis. Mak.* **2013**, *13*, 30. [CrossRef]
31. Afzal, N.; Sohn, S.; Abram, S.; Liu, H.; Kullo, I.J.; Arruda-Olson, A.M. Identifying peripheral arterial disease cases using natural language processing of clinical notes. In Proceedings of the 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Las Vegas, NV, USA, 24–27 February 2016; pp. 126–131.
32. Restrepo, N.A.; Farber-Eger, E.; Crawford, D.C. Searching in the Dark: Phenotyping Diabetic Retinopathy in a De-Identified Electronic Medical Record Sample of African Americans. *AMIA Summits Transl. Sci. Proc.* **2016**, *2016*, 221.
33. Li, D.; Simon, G.; Chute, C.G.; Pathak, J. Using association rule mining for phenotype extraction from electronic health records. *AMIA Summits Transl. Sci. Proc.* **2013**, *2013*, 142.
34. Doss, J.; Mo, H.; Carroll, R.J.; Crofford, L.J.; Denny, J.C. Phenome-wide association study of rheumatoid arthritis subgroups identifies association between seronegative disease and fibromyalgia. *Arthritis Rheumatol.* **2017**, *69*, 291–300. [CrossRef]
35. Aref-Eshghi, E.; Oake, J.; Godwin, M.; Aubrey-Bassler, K.; Duke, P.; Mahdavian, M.; Asghari, S. Identification of Dyslipidemic Patients Attending Primary Care Clinics Using Electronic Medical Record (EMR) Data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) Database. *J. Med. Syst.* **2017**, *41*, 45. [CrossRef]
36. Bobo, W.V.; Pathak, J.; Kremers, H.M.; Yawn, B.P.; Brue, S.M.; Stoppel, C.J.; Croarkin, P.E.; St Sauver, J.; Frye, M.A.; Rocca, W.A. An electronic health record driven algorithm to identify incident antidepressant medication users. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 785–791. [CrossRef] [PubMed]
37. Yu, S.; Liao, K.P.; Shaw, S.Y.; Gainer, V.S.; Churchill, S.E.; Szolovits, P.; Murphy, S.N.; Kohane, I.S.; Cai, T. Toward high-throughput phenotyping: Unbiased automated feature extraction and selection from knowledge sources. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 993–1000. [CrossRef] [PubMed]
38. Michalik, D.E.; Taylor, B.W.; Panepinto, J.A. Identification and validation of a sickle cell disease cohort within electronic health records. *Acad. Pediatr.* **2017**, *17*, 283–287. [CrossRef] [PubMed]
39. Connolly, B.; Miller, T.; Ni, Y.; Cohen, K.B.; Savova, G.; Dexheimer, J.W.; Pestian, J. Natural Language Processing–Overview and History. In *Pediatric Biomedical Informatics*; Springer: Berlin, Germany, 2016; pp. 203–230.

40. Nicholson, A.; Tate, A.R.; Koeling, R.; Cassell, J.A. What does validation of cases in electronic record databases mean? The potential contribution of free text. *Arthritis Rheumatol.* **2011**, *20*, 321–324. [CrossRef] [PubMed]
41. Rizzoli, P.; Loder, E.; Joshi, S. Validity of cluster headache diagnoses in an electronic health record data repository. *Headache J. Head Face Pain* **2016**, *56*, 1132–1136. [CrossRef]
42. Garg, R.; Dong, S.; Shah, S.; Jonnalagadda, S.R. A Bootstrap Machine Learning Approach to Identify Rare Disease Patients from Electronic Health Records. *arXiv* **2016**, arXiv:1609.01586.
43. Gundlapalli, A.V.; Redd, A.; Carter, M.; Divita, G.; Shen, S.; Palmer, M.; Samore, M.H. Validating a strategy for psychosocial phenotyping using a large corpus of clinical text. *J. Am. Med. Inform. Assoc.* **2013**, *20*, e355–e364. [CrossRef]
44. Spyns, P. Natural language processing. *Methods Inf. Med.* **1996**, *35*, 285–301.
45. Walsh, S.H. The clinician's perspective on electronic health records and how they can affect patient care. *BMJ* **2004**, *328*, 1184–1187. [CrossRef]
46. Earl, M.F. Information retrieval in biomedicine: Natural language processing for knowledge integration. *J. Med. Libr. Assoc. JMLA* **2010**, *98*, 190. [CrossRef]
47. Byrd, R.J.; Steinhubl, S.R.; Sun, J.; Ebadollahi, S.; Stewart, W.F. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *Int. J. Med. Inform.* **2014**, *83*, 983–992. [CrossRef] [PubMed]
48. Jha, A.K. The promise of electronic records: Around the corner or down the road? *JAMA* **2011**, *306*, 880–881. [CrossRef] [PubMed]
49. Wright, A.; McCoy, A.B.; Henkin, S.; Kale, A.; Sittig, D.F. Use of a support vector machine for categorizing free-text notes: assessment of accuracy across two institutions. *Int. J. Med. Inform.* **2013**, *20*, 887–890. [CrossRef] [PubMed]
50. Afzal, Z.; Engelkes, M.; Verhamme, K.; Janssens, H.M.; Sturkenboom, M.C.; Kors, J.A.; Schuemie, M.J. Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases. *Pharmacoepidemiol. Drug Saf.* **2013**, *22*, 826–833. [CrossRef] [PubMed]
51. Lin, C.; Karlson, E.W.; Canhao, H.; Miller, T.A.; Dligach, D.; Chen, P.J.; Perez, R.N.G.; Shen, Y.; Weinblatt, M.E.; Shadick, N.A.; et al. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. *PLoS ONE* **2013**, *8*, e69932. [CrossRef] [PubMed]
52. Cohen, K.B.; Glass, B.; Greiner, H.M.; Holland-Bouley, K.; Standridge, S.; Arya, R.; Faist, R.; Morita, D.; Mangano, F.; Connolly, B.; et al. Methodological Issues in Predicting Pediatric Epilepsy Surgery Candidates Through Natural Language Processing and Machine Learning. *Biomed. Inform. Insights* **2016**, *8*, 11. [CrossRef] [PubMed]
53. Kimia, A.A.; Savova, G.; Landschaft, A.; Harper, M.B. An introduction to natural language processing: How you can get more from those electronic notes you are generating. *Pediatric Emerg. Care* **2015**, *31*, 536–541. [CrossRef]
54. Nelson, R.E.; Butler, J.; LaFleur, J.; Knippenberg, K.C. Kamauu, A.W.; DuVall, S.L. Determining Multiple Sclerosis Phenotype from Electronic Medical Records. *J. Manag. Care Spec. Pharm.* **2016**, *22*, 1377–1382. [CrossRef]
55. Castro, V.M.; Minnier, J.; Murphy, S.N.; Kohane, I.; Churchill, S.E.; Gainer, V.; Cai, T.; Hoffnagle, A.G.; Dai, Y.; Block, S.; et al. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *Am. J. Psychiatry* **2015**, *172*, 363–372. [CrossRef]
56. Zeng, Q.T.; Goryachev, S.; Weiss, S.; Sordo, M.; Murphy, S.N.; Lazarus, R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. *BMC Med. Inform. Decis. Mak.* **2006**, *6*, 30. [CrossRef]
57. Harkema, H.; Dowling, J.N.; Thornblade, T.; Chapman, W.W. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J. Biomed. Inform.* **2009**, *42*, 839–851. [CrossRef] [PubMed]
58. Ludvigsson, J.F.; Pathak, J.; Murphy, S.; Durski, M.; Kirsch, P.S.; Chute, C.G.; Ryu, E.; Murray, J.A. Use of computerized algorithm to identify individuals in need of testing for celiac disease. *J. Am. Med. Inform. Assoc.* **2013**, *20*, e306–e310. [CrossRef] [PubMed]

59. Gundlapalli, A.V.; Divita, G.; Redd, A.; Carter, M.E.; Ko, D.; Rubin, M.; Samore, M.; Strymish, J.; Krein, S.; Gupta, K.; et al. Detecting the presence of an indwelling urinary catheter and urinary symptoms in hospitalized patients using natural language processing. *J. Biomed. Inform.* **2017**, *71*, S39–S45. [CrossRef] [PubMed]
60. Hanauer, D.A.; Gardner, M.; Sandberg, D.E. Unbiased identification of patients with disorders of sex development. *PLoS ONE* **2014**, *9*, e108702. [CrossRef]
61. Chary, M.; Parikh, S.; Manini, A.F.; Boyer, E.W.; Radeos, M. A Review of Natural Language Processing in Medical Education. *Western J. Emergency Med.* **2019**, *20*, 78. [CrossRef]
62. SNOMED, C. International Health Terminology Standards Development Organisation Web site, London, UK, 2014. Available online: http://www.snomed.org/ (accessed on 16 September 2017).
63. Fact, S.U. Metathesaurus® National Library of Medicine. *Metathesaurus [en línea]*. Available online: http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html (accessed on 8 May 2012).
64. Savova, G.K.; Masanz, J.J.; Ogren, P.V.; Zheng, J.; Sohn, S.; Kipper-Schuler, K.C.; Chute, C.G. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 507–513. [CrossRef]
65. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *J. Am. Med. Inform. Assoc.* **2001**, *2001*, 17–21.
66. Hristovski, D.; Friedman, C.; Rindflesch, T.C.; Peterlin, B. Exploiting semantic relations for literature-based discovery. AMIA annual symposium proceedings. *J. Am. Med. Inform. Assoc.* **2006**, *2006*, 349.
67. Denny, J.C.; Smithers, J.D.; Miller, R.A.; Spickard, A. "Understanding" medical school curriculum content using KnowledgeMap. *J. Am. Med. Inform. Assoc.* **2003**, *10*, 351–362. [CrossRef]
68. Lin, C.; Karlson, E.W.; Dligach, D.; Ramirez, M.P.; Miller, T.A.; Mo, H.; Braggs, N.S.; Cagan, A.; Gainer, V.; Denny, J.C.; et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *J. Am. Med. Inform. Assoc.* **2014**, *22*, e151–e161. [CrossRef]
69. Lingren, T.; Chen, P.; Bochenek, J.; Doshi-Velez, F.; Manning-Courtney, P.; Bickel, J.; Welchons, L.W.; Reinhold, J.; Bing, N.; Ni, Y.; et al. Electronic Health Record Based Algorithm to Identify Patients with Autism Spectrum Disorder. *PLoS ONE* **2016**, *11*, e0159621. [CrossRef] [PubMed]
70. Teixeira, P.L.; Wei, W.Q.; Cronin, R.M.; Mo, H.; VanHouten, J.P.; Carroll, R.J.; LaRose, E.; Bastarache, L.A.; Rosenbloom, S.T.; Edwards, T.L.; et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *J. Am. Med. Inform. Assoc.* **2016**, *24*, 162–171. [CrossRef] [PubMed]
71. Ananthakrishnan, A.N.; Cagan, A.; Cai, T.; Gainer, V.S.; Shaw, S.Y.; Savova, G.; Churchill, S.; Karlson, E.W.; Murphy, S.N.; Liao, K.P.; et al. Identification of nonresponse to treatment using narrative data in an electronic health record inflammatory bowel disease cohort. *Inflammatory Bowel Dis.* **2016**, *22*, 151–158. [CrossRef] [PubMed]
72. Ye, Y.; Tsui, F.; Wagner, M.; Espino, J.U.; Li, Q. Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 815–823. [CrossRef]
73. Luo, Y.; Sohani, A.R.; Hochberg, E.P.; Szolovits, P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 824–832. [CrossRef]
74. Hinz, E.R.M.; Bastarache, L.; Denny, J.C. A natural language processing algorithm to define a venous thromboembolism phenotype. *J. Am. Med. Inform. Assoc.*, **2013**, *2013*, 975.
75. Yadav, K.; Sarioglu, E.; Smith, M.; Choi, H.A. Automated outcome classification of emergency department computed tomography imaging reports. *Acad. Emerg. Med.* **2013**, *20*, 848–854. [CrossRef]
76. Liao, K.P.; Cai, T.; Savova, G.K.; Murphy, S.N.; Karlson, E.W.; Ananthakrishnan, A.N.; Gainer, V.S.; Shaw, S.Y.; Xia, Z.; Szolovits, P.; et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ* **2015**, *350*, h1885. [CrossRef]
77. Pineda, A.L.; Ye, Y.; Visweswaran, S.; Cooper, G.F.; Wagner, M.M.; Tsui, F.R. Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *J Biomed. Inf.* **2015**, *58*, 60–69. [CrossRef]
78. Chapman, W.W.; Bridewell, W.; Hanbury, P.; Cooper, G.F.; Buchanan, B.G. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. Biomed. Inf.* **2001**, *34*, 301–310. [CrossRef]

79. Chu, D. Clinical Feature Extraction from Emergency Department Reports for Biosurveillance. Master's Thesis, University of Pittsburgh, Pittsburgh, PA, USA, 2007.
80. Yu, S.; Cai, T. A short introduction to NILE. *arXiv* **2013**, arXiv:1311.6063.
81. Wagholikar, K.; Torii, M.; Jonnalagadda, S.; Liu, H. Feasibility of pooling annotated corpora for clinical concept extraction. *AMIA Summits Transl. Sci. Proc.* **2012**, *2012*, 38. [PubMed]
82. Xu, H.; Stenner, S.P.; Doan, S.; Johnson, K.B.; Waitman, L.R.; Denny, J.C. MedEx: A medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 19–24. [CrossRef] [PubMed]
83. de Quirós, F.G.B.; Otero, C.; Luna, D. Terminology Services: Standard Terminologies to Control Health Vocabulary. *Yearbook Med. Inf.* **2018**, *27*, 227–233.
84. Ma, S.; Huang, J. Penalized feature selection and classification in bioinformatics. *Brief. Bioinform.* **2008**, *9*, 392–403. [CrossRef] [PubMed]
85. Zhao, Z.; Morstatter, F.; Sharma, S.; Alelyani, S.; Anand, A.; Liu, H. Advancing feature selection research. *ASU Feature Sel. Repos.* **2010**, 1–28.
86. Garla, V.N.; Brandt, C. Ontology-guided feature engineering for clinical text classification. *J. Biomed. Inf.* **2012**, *45*, 992–998. [CrossRef]
87. Bejan, C.A.; Xia, F.; Vanderwende, L.; Wurfel, M.M.; Yetisgen-Yildiz, M. Pneumonia identification using statistical feature selection. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 817–823. [CrossRef]
88. Alzubi, R.; Ramzan, N.; Alzoubi, H.; Amira, A. A hybrid feature selection method for complex diseases SNPs. *IEEE Access* **2017**, *6*, 1292–1301. [CrossRef]
89. Greenwood, P.E.; Nikulin, M.S. *A Guide to Chi-Squared Testing*; John Wiley & Sons: Hoboken, NJ, USA, 1996.
90. Zhong, V.W.; Obeid, J.S.; Craig, J.B.; Pfaff, E.R.; Thomas, J.; Jaacks, L.M.; Beavers, D.P.; Carey, T.S.; Lawrence, J.M.; Dabelea, D.; et al. An efficient approach for surveillance of childhood diabetes by type derived from electronic health record data: The SEARCH for Diabetes in Youth Study. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 1060–1067. [CrossRef]
91. Fan, J.; Arruda-Olson, A.M.; Leibson, C.L.; Smith, C.; Liu, G.; Bailey, K.R.; Kullo, I.J. Billing code algorithms to identify cases of peripheral artery disease from administrative data. *J. Am. Med. Inform. Assoc.* **2013**, *20*, e349–e354. [CrossRef] [PubMed]
92. Oake, J.; Aref-Eshghi, E.; Godwin, M.; Collins, K.; Aubrey-Bassler, K.; Duke, P.; Mahdavian, M.; Asghari, S. Using electronic medical record to identify patients with dyslipidemia in primary care settings: International classification of disease code matters from one region to a national database. *Biomed. Inform. Insights* **2017**, *9*. [CrossRef]
93. Kagawa, R.; Kawazoe, Y.; Ida, Y.; Shinohara, E.; Tanaka, K.; Imai, T.; Ohe, K. Development of Type 2 Diabetes Mellitus Phenotyping Framework Using Expert Knowledge and Machine Learning Approach. *J. Diabetes Sci. Technol.* **2017**, *11*, 791–799. [CrossRef] [PubMed]
94. Wing, K.; Bhaskaran, K.; Smeeth, L.; van Staa, T.P.; Klungel, O.H.; Reynolds, R.F.; Douglas, I. Optimising case detection within UK electronic health records: Use of multiple linked databases for detecting liver injury. *BMJ Open* **2016**, *6*, e012102. [CrossRef] [PubMed]
95. Zhou, S.M.; Fernandez-Gutierrez, F.; Kennedy, J.; Cooksey, R.; Atkinson, M.; Denaxas, S.; Siebert, S.; Dixon, W.G.; O'Neill, T.W.; Choy, E.; et al. Defining disease phenotypes in primary care electronic health records by a machine learning approach: A case study in identifying rheumatoid arthritis. *PLoS ONE* **2016**, *11*, e0154515. [CrossRef] [PubMed]
96. Anderson, A.E.; Kerr, W.T.; Thames, A.; Li, T.; Xiao, J.; Cohen, M.S. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. *J. Biomed. Inform.* **2016**, *60*, 162–168. [CrossRef] [PubMed]
97. Wu, S.T.; Sohn, S.; Ravikumar, K.; Wagholikar, K.; Jonnalagadda, S.R.; Liu, H.; Juhn, Y.J. Automated chart review for asthma cohort identification using natural language processing: An exploratory study. *Ann. Allergy Asthma Immunol.* **2013**, *111*, 364–369. [CrossRef] [PubMed]
98. Mowery, D.L.; Chapman, B.E.; Conway, M.; South, B.R.; Madden, E.; Keyhani, S.; Chapman, W.W. Extracting a stroke phenotype risk factor from Veteran Health Administration clinical reports: An information content analysis. *J. Biomed. Semant.* **2016**, *7*, 26. [CrossRef]
99. DeLisle, S.; Kim, B.; Deepak, J.; Siddiqui, T.; Gundlapalli, A.; Samore, M.; D'Avolio, L. Using the electronic medical record to identify community-acquired pneumonia: Toward a replicable automated strategy. *PLoS ONE* **2013**, *8*, e70944. [CrossRef]

100. Valkhoff, V.E.; Coloma, P.M.; Masclee, G.M.; Gini, R.; Innocenti, F.; Lapi, F.; Molokhia, M.; Mosseveld, M.; Nielsson, M.S.; Schuemie, M.; et al. Validation study in four health-care databases: Upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. *J. Clin. Epidemiol.* **2014**, *67*, 921–931. [CrossRef]
101. Liu, H.; Bielinski, S.J.; Sohn, S.; Murphy, S.; Wagholikar, K.B.; Jonnalagadda, S.R.; Ravikumar, K.; Wu, S.T.; Kullo, I.J.; Chute, C.G. An information extraction framework for cohort identification using electronic health records. *AMIA Summits Trans. Sci. Proc.* **2013**, *2013*, 149.
102. Mo, H.; Thompson, W.K.; Rasmussen, L.V.; Pacheco, J.A.; Jiang, G.; Kiefer, R.; Zhu, Q.; Xu, J.; Montague, E.; Carrell, D.S.; et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 1220–1230. [CrossRef] [PubMed]
103. Xi, N.; Wallace, R.; Agarwal, G.; Chan, D.; Gershon, A.; Gupta, S. Identifying patients with asthma in primary care electronic medical record systems. *Can. Fam. Physician* **2015**, *61*, e474–e483.
104. Roch, A.M.; Mehrabi, S.; Krishnan, A.; Schmidt, H.E.; Kesterson, J.; Beesley, C.; Dexter, P.R.; Palakal, M.; Schmidt, C.M. Automated pancreatic cyst screening using natural language processing: A new tool in the early detection of pancreatic cancer. *HPB* **2015**, *17*, 447–453. [CrossRef] [PubMed]
105. Thomas, A.A.; Zheng, C.; Jung, H.; Chang, A.; Kim, B.; Gelfond, J.; Slezak, J.; Porter, K.; Jacobsen, S.J.; Chien, G.W. Extracting data from electronic medical records: Validation of a natural language processing program to assess prostate biopsy results. *World J. Urol.* **2014**, *32*, 99–103. [CrossRef]
106. Jackson, K.L.; Mbagwu, M.; Pacheco, J.A.; Baldridge, A.S.; Viox, D.J.; Linneman, J.G.; Shukla, S.K.; Peissig, P.L.; Borthwick, K.M.; Carrell, D.A.; et al. Performance of an electronic health record-based phenotype algorithm to identify community associated methicillin-resistant Staphylococcus aureus cases and controls for genetic association studies. *BMC Infect. Dis.* **2016**, *16*, 684. [CrossRef]
107. Safarova, M.S.; Liu, H.; Kullo, I.J. Rapid identification of familial hypercholesterolemia from electronic health records: The SEARCH study. *J. Clin. Lipidol.* **2016**, *10*, 1230–1239. [CrossRef]
108. Chartrand, S.; Swigris, J.J.; Stanchev, L.; Lee, J.S.; Brown, K.K.; Fischer, A. Clinical features and natural history of interstitial pneumonia with autoimmune features: A single center experience. *Respir. Med.* **2016**, *119*, 150–154. [CrossRef]
109. Alpaydin, E. *Introduction to Machine Learning*; MIT Press: Cambridge, UK, 2014.
110. Henriksson, A. Semantic Spaces of Clinical Text: Leveraging Distributional Semantics for Natural Language Processing of Electronic Health Records. Ph.D. Thesis, Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden, 2013.
111. Alzoubi, H.; Ramzan, N.; Alzubi, R.; Mesbahi, E. An Automated System for Identifying Alcohol Use Status from Clinical Text. In Proceedings of the 2018 IEEE International Conference on Computing, Southend, UK, 16–17 Auguest 2018; pp. 41–46.
112. Huda, S.; Abawajy, J.; Alazab, M.; Abdollalihian, M.; Islam, R.; Yearwood, J. Hybrids of support vector machine wrapper and filter based framework for malware detection. *Future Gener. Comp. Sys.* **2016**, *55*, 376–390. [CrossRef]
113. Lasko, T.A.; Denny, J.C.; Levy, M.A. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS ONE* **2013**, *8*, e66341. [CrossRef]
114. Lipton, Z.C.; Kale, D.C.; Elkan, C.; Wetzel, R. Learning to diagnose with LSTM recurrent neural networks. *arXiv* **2015**, arXiv:1511.03677.
115. Gehrmann, S.; Dernoncourt, F.; Li, Y.; Carlson, E.T.; Wu, J.T.; Welt, J.; Foote, J.J.; Moseley, E.T.; Grant, D.W.; Tyler, P.D.; et al. Comparing Rule-Based and Deep Learning Models for Patient Phenotyping. *arXiv* **2017**, arXiv:1703.08705.
116. Kale, D.C.; Che, Z.; Bahadori, M.T.; Li, W.; Liu, Y.; Wetzel, R. Causal phenotype discovery via deep networks. AMIA Annual Symposium Proceedings. *J. Am. Med. Inform. Assoc.* **2015**, *2015*, 677.
117. Zheng, C.; Rashid, N.; Wu, Y.L.; Koblick, R.; Lin, A.T.; Levy, G.D.; Cheetham, T.C. Using natural language processing and machine learning to identify gout flares from electronic clinical notes. *Arthritis Care Res.* **2014**, *66*, 1740–1748. [CrossRef]
118. Ho, J.C.; Ghosh, J.; Sun, J. Extracting phenotypes from patient claim records using nonnegative tensor factorization. International Conference on Brain Informatics and Health. *J. Biomed. Inform.* **2014**, *52*, 199–211. [CrossRef]

119. Joshi, S.; Gunasekar, S.; Sontag, D.; Joydeep, G. Identifiable phenotyping using constrained non-negative matrix factorization. In Proceedings of the Machine Learning for Healthcare Conference, Los Angeles, CA, USA, 19–20 August 2016; pp. 17–41.
120. Gunasekar, S.; Ho, J.C.; Ghosh, J.; Kreml, S.; Kho, A.N.; Denny, J.C.; Malin, B.A.; Sun, J. Phenotyping using Structured Collective Matrix Factorization of Multi–source EHR Data. *arXiv* **2016**, arXiv:1609.04466.
121. Elmasry, W.; Akbulut, A.; Zaim, A.H. Deep learning approaches for predictive masquerade detection. *Secur.Commun. Net.* **2018**, *2018*. 1–24. [CrossRef]
122. Vazquez Guillamet, R.; Ursu, O.; Iwamoto, G.; Moseley, P.L.; Oprea, T. Chronic obstructive pulmonary disease phenotypes using cluster analysis of electronic medical records. *Health Inf. J.* **2016**, 394–409. [CrossRef]
123. Ho, J.C.; Ghosh, J.; Steinhubl, S.R.; Stewart, W.F.; Denny, J.C.; Malin, B.A.; Sun, J. Limestone: High-throughput candidate phenotype generation via tensor factorization. *J. Biomed. Inf.* **2014**, *52*, 199–211. [CrossRef]
124. Ho, J.C.; Ghosh, J.; Sun, J. 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, New York, NY, USA, 24–27 August 2014; pp. 115–124.
125. Wang, Y.; Chen, R.; Ghosh, J.; Denny, J.C.; Kho, A.; Chen, Y.; Malin, B.A.; Sun, J. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; pp. 1265–1274.
126. Schuler, A.; Liu, V.; Wan, J.; Callahan, A.; Udell, M.; Stark, D.E.; Shah, N.H. Discovering patient phenotypes using generalized low rank models. *Biocomputing* **2016**, *21*, 144–155.
127. Nguyen, D.H.; Patrick, J.D. Supervised machine learning and active learning in classification of radiology reports. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 893–901. [CrossRef] [PubMed]
128. Reddy, C.K.; Aggarwal, C.C. *Healthcare Data Analytics*; Chapman and Hall/CRC: London, UK, 2015.