OPEN ACCESS

A review of approaches to identifying patient phenotype cohorts using electronic health records

Chaitanya Shivade,[1] Preethi Raghavan,[1] Eric Fosler-Lussier,[1] Peter J Embi,[2] Noemie Elhadad,[3] Stephen B Johnson,[4] Albert M Lai[2]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001935).

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA
[2]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA
[3]Department of Biomedical Informatics, Columbia University, New York, New York, USA
[4]Center for Healthcare Informatics and Policy, Weill Cornell Medical College, New York, New York, USA

Correspondence to
Chaitanya Shivade, Department of Computer Science and Engineering, The Ohio State University, 395 Dreese Laboratories, 2015 Neil Avenue, Columbus, OH 43210, USA; shivade@cse.ohio-state.edu

Received 15 April 2013
Revised 18 October 2013
Accepted 25 October 2013
Published Online First 7 November 2013

To cite: Shivade C, Raghavan P, Fosler-Lussier E, *et al*. *J Am Med Inform Assoc* 2014;**21**:221–230.

ABSTRACT

Objective To summarize literature describing approaches aimed at automatically identifying patients with a common phenotype.

Materials and methods We performed a review of studies describing systems or reporting techniques developed for identifying cohorts of patients with specific phenotypes. Every full text article published in (1) *Journal of American Medical Informatics Association*, (2) *Journal of Biomedical Informatics*, (3) *Proceedings of the Annual American Medical Informatics Association Symposium*, and (4) *Proceedings of Clinical Research Informatics Conference* within the past 3 years was assessed for inclusion in the review. Only articles using automated techniques were included.

Results Ninety-seven articles met our inclusion criteria. Forty-six used natural language processing (NLP)-based techniques, 24 described rule-based systems, 41 used statistical analyses, data mining, or machine learning techniques, while 22 described hybrid systems. Nine articles described the architecture of large-scale systems developed for determining cohort eligibility of patients.

Discussion We observe that there is a rise in the number of studies associated with cohort identification using electronic medical records. Statistical analyses or machine learning, followed by NLP techniques, are gaining popularity over the years in comparison with rule-based systems.

Conclusions There are a variety of approaches for classifying patients into a particular phenotype. Different techniques and data sources are used, and good performance is reported on datasets at respective institutions. However, no system makes comprehensive use of electronic medical records addressing all of their known weaknesses.

INTRODUCTION

The identification of patients who satisfy predefined criteria from a large population in an institution has numerous use cases, including clinical trial recruitment, outcome prediction, survival analysis, and other kinds of retrospective studies.[1 2] However, the process of distinguishing these patients on the basis of their patient records can be extremely time-consuming and challenging depending on the complexity of the criteria. This is because the data matching these criteria are buried within multiple documents and across multiple data points in the electronic health record (EHR) of a patient. Some data, such as laboratory results, medications, and diagnoses, have a structured format. Clinicians provide important additional observations in unstructured text, such as radiology reports, progress notes, discharge summaries, and other clinical narratives. In addition to the open challenges inherent in parsing the often-complex clinical narrative, the presence of ungrammatical text, local dialectal phrases, abbreviations and misspellings, boilerplate and template text make the task of processing these documents even harder. Furthermore, the aggregate information in the structured and unstructured parts of the EHR may be implicit or explicit and may require reconciliation strategies. The ability to extract meaningful pieces of information from the EHR and consolidate them into a coherent structure would provide great value for automatically identifying patient cohorts that satisfy complex criteria.

A large number of studies have been published describing automated phenotyping techniques employed by medical organizations across the USA. Owing to the sensitive nature of patient data, administrative roadblocks, and collaboration overheads, most institutions have developed their own systems. There are efforts to extend techniques developed at one site across multiple sites. However, there are no established standard tools available that an institution can pick up and start using without significant challenges. There is little clarity regarding the nature of a phenotyping solution that will work at any given institution. This review aims to develop an understanding of the techniques used by different organizations, the processes followed by them, and, ultimately, the best practices for building a successful phenotyping solution.

We present a review of the state-of-the art literature on patient cohort identification. The objectives were to (1) compile the studied phenotypes, (2) inspect the data sources commonly used, (3) study the different approaches that have been successful in the past, and (4) portray the role of large-scale systems in the development of cohort identification.

METHOD

Design

EHR-based phenotyping has not been well defined in literature and its meaning is thus wide ranging. We limited our scope of phenotyping to only those studies that explicitly investigated identification of patient cohorts. We realized that many important articles were a part of conference proceedings, which are not indexed by Medline. Attempts to search PubMed using sophisticated queries with a combination of keywords and MeSH terms appeared to miss a large number of relevant studies. These studies essentially performed the same task of cohort identification but tended to

focus on different issues such as evaluation of predictive models, end-to-end system descriptions, comparison of analysis techniques, exploration of data sources, and evaluation of technologies. (Refer to online supplementary appendix 1 for a sample query and its analysis.) To address this limitation, we manually reviewed all the issues of (1) *Journal of American Medical Informatics Association* (JAMIA), (2) *Journal of Biomedical Informatics* (JBI), (3) *Proceedings of the Annual American Medical Informatics Association Symposium* (AMIA), and (4) *Proceedings of the AMIA Clinical Research Informatics Conference* (CRI) from the years 2010–2012. Our authorship team deemed these to be venues that would probably have the highest density of publications on electronic phenotyping. The intent of this review is to demonstrate the types of work being conducted rather than being a comprehensive list of all work in the field of EHR-based phenotyping.

The analysis used was a three-step process. The first step involved filtering relevant articles by reading the abstract and title of the articles. We identified 76 articles (26 from JAMIA, 11 from JBI, 27 from AMIA, and 12 from CRI) in this step. In the second step, we reviewed the titles and abstracts of all the references that satisfied our criteria and were cited by these articles. This resulted in an additional 53 articles after removal of duplicates. Finally, we read the full text of these articles and discarded 32 articles, resulting in a final set of 97 articles. The entire process was carried out by two authors, CS and PR, with a κ statistic of 0.78. Conflicts were resolved by a third author, AML. The final set of studies was chosen on the basis of discussion.

Inclusion and exclusion criteria

We included studies that: (1) described identification of patients with particular diagnoses or a medical condition; (2) were perspectives or characterizations of clinical trial recruitment solutions; (3) described novel techniques, compared different methods, or investigated diverse data sources for cohort identification. We followed the broad but commonly accepted definition of a phenotype as being the observable characteristic of an organism (in our case, the patient). Thus, identification of patients based on the stage of cancer, drug side effects, smoking status, infections, or response to therapy were all included. We discarded studies that (1) used only manual techniques, (2) relied only on data sources that were not EHRs, or (3) were not related to identifying patient cohorts. Many studies describe tools and techniques that can be potentially used for phenotyping but focus on proposing or evaluating a new methodology. Such studies were not considered in our review (see online supplementary appendix 2 for illustrative examples and reasons for their inclusion or exclusion in the review). Although the development of techniques for representing eligibility criteria to ease automatic cohort identification is an interesting and closely related topic, we did not include any articles investigating this issue (figure 1).

Phenotype under consideration

Studies in our review addressed different diagnoses as the phenotype of interest. A large number of papers focused on diabetes, cancer, heart failure,[3–7] rheumatoid arthritis,[8–12] or cataract.[13–16] A few studies described generic methods that could be applied to multiple diagnoses. Several papers addressed identification of adverse drug events.[17–22] Some studies also analyzed genomic data of a predetermined phenotype to gain insight into other phenotypic properties of the same cohort. Table 1 shows the top 10 phenotypes of interest. Although cancer has been

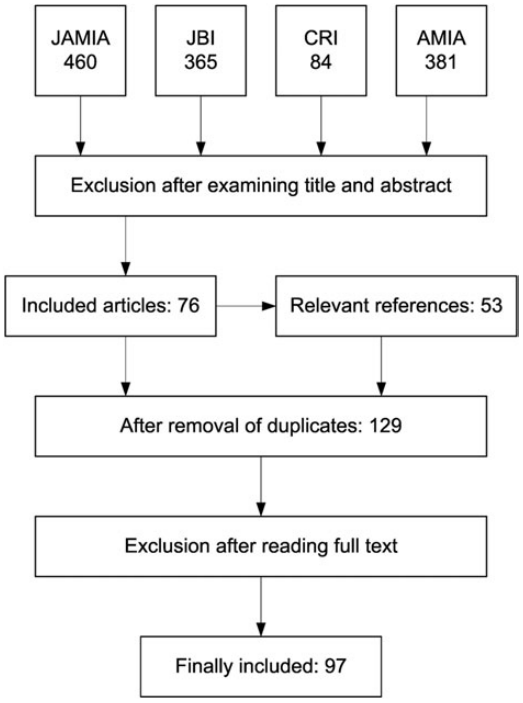

Figure 1 Flow chart of study inclusion.

reported as a generic phenotype, there were 12 types of cancer (with breast cancer being most prevalent) studied in seven articles.[23–29] Congestive heart failure and heart failure were counted as the same phenotype. Similarly, hypertension and resistant hypertension were considered to be one phenotype.[14 15 30] Some articles also considered other observable characteristics such as smoking status[31–34] and obesity[30 35] among patients. Pneumonia[36–39] and a variety of other infectious diseases[40–48] were studied.

Data sources

Previous studies[8 49 50] have concluded that use of International Classification of Diseases, Ninth Revision (ICD-9) codes is not sufficient and have encouraged the use of additional sources of data or analysis techniques for identifying patient cohorts. Similarly, patient history data are also considered to be insufficient.[51] Studies included in this review use diverse data sources, such as patient demographics, medications, laboratory reports, vital signs, clinical data, diagnoses, treatment, clinical notes, or even genomic data. We observed that diagnosis codes and

Table 1 Top 10 phenotypes of interest

Phenotype of interest	Number of studies
Cancer	26
Diabetes	23
Heart failure	5
Rheumatoid arthritis	5
Cataract	4
Drug side effect	4
Pneumonia	4
Asthma	3
Peripheral arterial disease	3
Hypertension	3

patient demographics were commonly used in rule-based systems. Treatment data were included based on use of current procedural terminology (CPT) codes and billing codes or as explicit mentions in the text. Clinical notes here account for any textual document in the EHR, such as discharge summaries, progress reports, and pathology reports. Clinical data generally refer to variables that are specific clinical findings (eg, presence of a device, consumption of alcohol) or an aggregation of variables from multiple sources (eg, congestive heart failure, uncompensated or ejection fraction (EF) <25%, therapy period). Popular categories among other data sources include imaging data,[25 52] insurance claim data, drug characterization databases,[21] scientific articles from resources such as PubMed,[53] and public health statistics. Table 2 summarizes data sources used across the phenotypes in table 1. We found that multiple sources of structured data such as diagnoses, medications and laboratory reports were often used together. Some studies used only clinical notes but they were most commonly used with or compared against diagnoses information.

Performance of phenotyping techniques is reported using different metrics. While specificity, sensitivity and positive predictive value (PPV) are most commonly reported, some studies report area under the receiver operating characteristics curve or F1-measure. Some studies also report p value significance or an agreement statistic between the proposed automated technique and human annotations. While many studies (33%) have separate training and testing datasets, some studies (18%) report cross-validation results. The number of folds used for cross-validation also varies across studies. Studies are conducted on datasets at individual institutions, making it difficult to compare them.

Manually reviewed data (76%) are used as the 'ground' truth by most of the studies reporting performance on data. Some studies (19%) compare the performance of the proposed automated technique with ICD-9 Clinical Modification (ICD-9-CM) codes from the EHR, while the remaining (5%) studies systematically calculate the ground truth using other variables from the dataset. Shared tasks and challenges are the only avenues where different methods are directly comparable using a common dataset and evaluation strategy. Size of dataset is reported in terms of different units such as number of patients, number of documents, number of notes, and number of samples, with number of patients being the most common unit (see online supplementary appendix 3 for distribution of dataset size across studies included in this review).

RESULTS

In this section, we discuss the tools and techniques used across different studies considered in our review (see online supplementary appendix 4 for the distribution of techniques used across the top 10 phenotypes). These studies describe the process of obtaining a phenotype, as well as refining it. We elaborate on the ones that have been published recently and have a higher number of citations. Other relevant studies are cited in appropriate sections, but are not discussed in detail.

Rule-based systems

We describe algorithms that deduce phenotypes of patients, by applying logical constraints (rules) to discrete values (eg, hemoglobin <10 AND age >60) extracted from an EHR as rule-based systems. A typical rule-based system applies these constraints to multiple values in a single step or a sequence of steps. In this section, we discuss rule-based systems, considering the way the rules were generated.

Rules based on clinical judgment

Most of the systems derived rules using clinical judgment of physicians or expert opinions. Nguyen *et al*[54] built a symbolic rule-based classification system for identifying lung cancer stages based on text occurrences. They used a tool called MEDTEX, which comprises modules for mapping free text to Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) terms, negation finder, and possibility identification. The performance of their system was comparable to that of support vector machine (SVM)-based text classification systems. They argue that highly discriminatory rules could be developed using a small fraction of the training data, in contrast with machine learning approaches, which also increases annotation overhead. Schmiedeskamp *et al*[47] developed empirical rules based on ICD-9-CM codes and laboratory and medication data to identify patients with nosocomial *Clostridium difficile* infection. They concluded that using data from a variety of sources yielded the best rule. Penberthy *et al*[23] created a two-step screening process for identifying patients eligible for cancer clinical trials. The first step involved analyzing commonly collected discrete data, followed by screening of text-based health level 7 (HL7) messages representing dictations from surgical pathology. The authors reported performance in terms of time and effort saved by using the software in comparison with manual efforts.

Rules based on healthcare guidelines

Many studies converted guidelines or recommendations from health organizations for specific diagnoses into rules. Kho *et al*[55] developed an algorithm to identify type 2 diabetes cases and controls using commonly collected EHR data across multiple institutions. Existing clinical diagnostic criteria established by the American Diabetes Association were used to develop an algorithm based on diagnostic codes, medications and laboratory test results. Small changes were made to ensure that the algorithm was portable across multiple institutions. The authors attribute their algorithm's PPV to multiple iterations and manual chart review. Klompas *et al*[42] used the clinical surveillance definition of acute hepatitis B published by the Centers for Disease Control and Prevention to create simple rules for identifying patients with this disease. Algorithms were tweaked on the basis of performance with training data to finally yield a single rule that achieves a very good performance. Trick *et al*[44] developed a system to detect bloodstream infections in patients on the basis of simple rules corresponding to National Nosocomical Infection Surveillance (NNIS) definitions to categorize blood isolates. Although their system was reliable, they found that the performance of the rules varied between the two sites of their study. Mathias *et al*[1] used guidelines from the American Cancer Society, American College of Obstetrics and Gynecology, and US Preventive Services Task Force to identify women eligible for cervical cancer screening. Their findings indicate that, although their method was not efficient in general, they could identify certain rare cases that would have been missed otherwise.

Refinement of previous rules

Some studies analyzed errors of rules developed in previous studies and modified them to generate better results. Hebert *et al*[56] in the past developed a system that used a variety of documents from insurance claims to identify diabetic patients. They used a rule set created by looking at previous studies that used similar data, and then extended it. Wright *et al*[57] used a previously created database of medication–problem and

Table 2 Summary of number of studies using different data sources across the top 10 phenotypes

	Demographics	Medications	Lab reports	Vitals	Clinical	Diagnosis	Treatment	Notes	Genomic	Other
Cancer	12	0	6	1	9	10	4	9	4	5
Diabetes	11	10	14	4	7	14	7	8	3	7
Heart failure	1	1	2	0	0	2	0	1	0	0
Rheumatoid arthritis	1	3	2	0	0	5	0	4	0	0
Cataract	3	1	1	1	1	3	2	2	0	3
Drug side effect	1	0	1	0	0	1	0	1	0	3
Pneumonia	0	0	0	0	0	0	0	4	0	0
Asthma	2	1	1	0	1	3	0	2	0	0
Peripheral arterial disease	2	1	1	1	1	2	2	1	0	2
Hypertension	1	0	0	0	0	0	0	2	0	1

laboratory–problem associations. Rules were added by reviewing medical textbooks and online clinical resources. The final set of rules was created after multiple iterations of review by physicians. Their method outperformed the problem list for all 17 conditions and billing codes for 12 of 17 conditions in the test dataset.

Automatically generated rules

Very few studies generated rules in an automated fashion. Li *et al*[58] followed the Quality Data Model in representing variables to identify cases for coronary artery disease and diabetes. They further used the RETE algorithm implemented in the JBoss Drools rules management system to automatically extract rules. Lee *et al*[59] used feature-selection algorithms on variables found to be relevant for predicting emergency room re-admission using past studies. These discriminatory variables were then analyzed by a Discriminatory Analysis Model to identify a classification rule.

Rule-based systems are easy to interpret, fast to implement, and give good results on limited datasets. While the evaluation of systems in the Informatics for Integrating Biology and the Bedside (i2b2) shared task on obesity identification[35] showed that the best predictions came from systems that processed text through rules, the smoking detection challenge[33] showed that rule-based systems performed just as well as other approaches. Many studies develop rules based on clinical expertise and knowledge of physicians. The rationale behind these rules is often not explained in detail. Following guidelines and recommendations of health organizations seems to be a good approach and gives promising results. However, it would be interesting to see if the details of these rules—for example, cut-off values for discretization—are indeed reflected by the data. As stated above, there are very few studies that explore automated rule mining.

Natural language processing

Clinical notes entered by physicians are valuable sources of patient information.[60 61] Clinical notes are often the only source of information from which to infer important phenotypic characteristics, which cannot be obtained from other data sources. Natural language processing (NLP) techniques have been successfully applied in other domains to perform a variety of tasks. However, misspellings, redundancy, ambiguity, and other characteristics of clinical text make the task challenging. Hence there is a need to systematically adapt NLP techniques from other domains to the realm of clinical text. In this section, we discuss studies that have made use of unstructured data for patient cohort identification, considering different approaches taken for a particular task (see online supplementary appendix 5 for a summary of different types of notes used across the top 10 phenotypes).

Term extraction

Most studies mapped textual elements to create Unified Medical Language System (UMLS) concepts for standardization. Bejan *et al*[39] used MetaMap[62] to extract UMLS concepts and all possible unigrams and bigrams from clinical reports and further ranked their relevance to pneumonia identification using statistical feature selection techniques. They found that the best performance could be achieved by considering only the top 25% features. McCowan *et al*[63] describe software for grouping patients based on cancer staging. They developed their own software, which follows a four-step process for normalizing text, mapping it to UMLS terms, and detecting and handling negations. Liao *et al*[9] compared the performance of an NLP system with codified EHR data for identifying subjects with rheumatoid arthritis. They used HITex[34] for extracting clinical information from narrative text. HITex, based on the GATE framework among other sophisticated modules, has a noun phrase finder and a UMLS concept mapper. Carroll *et al*[10] used Knowledge Map Concept Identifier, which processes clinical notes and returns UMLS concept identifiers handling negations. These concepts are used to identify cases of rheumatoid arthritis. Lehman *et al*[64] used their own software to map textual elements into SNOMED-CT concepts and further perform risk stratification of patients in the intensive care unit.

Use of keywords

Some studies use clinical knowledge or heuristics to identify keywords specific to the phenotype of interest. Sohn *et al*,[19] along with a sophisticated NLP pipeline, also make use of relevant keywords to identify cases of drug side effects. They comment that, although their rules could extract clearly stated side effects, improved semantic approaches are required to handle complex side effect descriptions. Sohn and Savova[31] improved the performance of their smoking detection module by adding selected temporal resolution keywords and dates as features. Simple regular expressions are also used to extract relevant parts of text from clinical notes. CUIMANDREef[3] is a simple regular expression-based system that extracts EF from free text echocardiogram reports to classify heart failures patients with abnormal EF. Friedlin *et al*[65] discuss the use of REX, a regular expression- and rule-based system, for extracting

relevant concepts in the context of identifying patients with pancreatic cancer.

After the extraction of relevant terms and concepts from free text reports, either a rule- or machine learning-based model is used to classify patients into cohorts. Rule-based systems are described in the previous section, while machine learning approaches are discussed in the next section.

Semantic web technologies

Few studies discuss the use of semantic web technologies for phenotyping. Cui *et al*[66] developed EpiDEA, an ontology-based epilepsy cohort identification system. While EpiDEA operates on unstructured text from discharge summaries using the full seven-stage clinical Text Analysis and Knowledge Extraction System (cTAKES)[67] pipeline, it uses a simple approach that extracts attribute–value pairs to work with semi-structured text. Pathak *et al*[68] describe the architecture of a system that leverages semantic web technologies for phenotyping and illustrate it through a case study[69] of type 2 diabetes.

Conway *et al*[15] surveyed the authors of 14 different phenotyping algorithms in their study analyzing the heterogeneity of such algorithms. The responders identified generation of NLP content as the most difficult aspects of algorithm construction. The studies in our review reflect this too. We observe that there is a lack of comprehensive NLP tools and datasets that exist as freeware for studying clinical data. MetaMap, cTAKES, and MedLEE are the most widely used tools. MetaMap is useful, but is limited to the identification and mapping of concepts from clinical text to the UMLS Metathesaurus. Although MedLEE is the oldest NLP tool[70] for such studies, it is only available through licensing. Furthermore, while MedLEE identifies many modifiers along with named entities, such as negation and history, it does not extract relations among entities.[71] cTAKES is open-source and free, but has only recently come out of the incubation stage and into a top-level project with the Apache Foundation. Most of the studies analyzed in this review developed their own NLP software for a specific task, with very few of the authors making them freely available. Most of the studies do not report handling of assertions for cases such as negations or speculative language. When use of negation is reported, Negex[72] is found to be a widely used tool.

Machine learning and statistical analysis

Machine learning methods learn patterns from data, in the form of parameters, to distinguish between certain predefined classes of interest. The field of biomedical informatics has embraced these methods for a variety of tasks. We discuss in this section studies that used these algorithms for patient cohort identification.

Comparison of popular algorithms

Many studies report the comparison of machine learning models for the same task. Sesen *et al*[73] compared Bayesian Networks and Naïve Bayes algorithms for predicting survival and recommending treatment in patients with lung cancer. They also used Markov Chain Monte Carlo Model Composition and the K2 greedy algorithm for structure learning. Their results on the English Lung Cancer Database show that the Naïve Bayes approach outperforms Bayesian Networks in predicting survival, while the converse is true for recommended treatments. Kawaler *et al*[74] also concluded that Naïve Bayes, along with random forests and SVM, are the best learners for predicting post-hospitalization venous thromboembolism risk. They also experimented with meta-learning approaches, such as bagging and boosting, without significant improvements. iDiagnosis[53] is a novel software combining knowledge extracted from PubMed and EHRs for predicting pancreatic cancer. This study showed that a weighted Bayesian Network Inference model outperformed models built using K-nearest neighbor or SVM. The authors comment that this may be due to the small feature set used in their models.

Decision tree-based algorithms

Few studies have used decision tree-based models. Van den Bulcke *et al*[75] found that regression models performed better than C4.5 decision trees in identifying cases of modeling of medium-chain acyl-CoA dehydrogenase deficiency. They observed that decision trees were less robust, with changes in variable selection choices. Mani *et al*[25] explored classification and regression trees and random forest models with three other types of machine learning techniques—linear classifiers, kernel-based methods, and rule learners—for predicting response of breast tumors to neoadjuvant chemotherapy. They found that tree-based models performed well on imaging data, while regression models were better for other sources of data. They followed a similar approach for type 2 diabetes,[76] where tree-based models outperformed other algorithms.

Other approaches

Other approaches have also been explored. Tatari *et al*[26] used multi-agent fuzzy systems to identify patients with a high risk of breast cancer. Lehman *et al*[64] carried out topic modeling, using hierarchical Dirichlet processes on unstructured notes of patients in an intensive care unit to group them into interesting categories such as 'on ventilator', 'post-cardiac surgery', and 'trauma', among others. Kim *et al*[77] used a semi-supervised graph propagation algorithm as an integrated framework that utilizes multilevel genomic data for prediction of clinical outcomes in brain cancer and ovarian cancer. Oberg *et al*[78] tried an interesting experiment exploring use of Google Search Appliance (GSA) for patient cohort discovery. They concluded that GSA did not have any significant advantage over traditional database querying.

Statistical methods

Traditional statistical techniques have also been used. Wang *et al*[79] compared three parametric models—Weibull exponential, log-logistic, and log-normal—for performing survival analysis predicting benefit of adjuvant chemoradiotherapy in gallbladder cancer. The log-normal model gave the best Akaike Information Criterion and exhibited better goodness of fit. Kim *et al*[80] report discriminative accuracy and the concordance statistics for a risk index that they developed based on a regression model predicting cancer risk among patients. Fine *et al*[48] developed a decision model incorporating heterogeneous data sources for identifying infants with pertussis. They conducted univariate analysis using χ^2 tests followed by multivariate regression analysis to develop a final logistic regression model. PheWAS[81] is an approach for determining which phenotypes are associated with a given genotype. The authors collected genotypic data at five single-nucleotide polymorphisms (SNPs) with previously reported disease associations. Their algorithm replicated four of the seven SNP–disease associations with statistically significant p values.

Although statistical analyses have been commonly used, the use of machine learning techniques has been scarce in the past. However, with the growing size of datasets, there are many studies exploring these methods. Although a variety of

algorithms have been used, we observe that logistic regression and SVMs are popular choices.

Hybrid approaches

Hybrid systems make use of both rule-based and statistical machine learning or NLP approaches. Such a system can be built in a number of ways. Typically a machine learning algorithm is fed with features that are a collection of rules and NLP-extracted attributes from clinical text.

Liu *et al*[32] describe changes made to the smoking detection module in cTAKES, developed at Mayo Clinic to yield a significantly better classifier in terms of the F-measure on a dataset obtained from Vanderbilt University. They filtered notes, added new annotated data for training the machine learning classifier (SVM with a radial basis function kernel), and added rules to the rule-based classifier. An approach using heterogeneous sources of data in an EHR was developed by Peissig *et al*[16] to identify cataract cases. The strategy was a combination of conventional data mining, NLP of free text documents, and optical character recognition of scanned clinical images. The multimodal approach was found to increase the case identification by a factor of three compared with single-mode approaches. Pathak *et al*[68] describe a data warehousing architecture along with a case study of type 2 diabetes[69] using semantic web technologies to enable better patient cohort discovery.

Xu *et al*[82] developed an algorithm combining machine learning and NLP to detect patients with colorectal cancer (CRC). Their two-step method extracted CRC concepts from clinical notes followed by determination of CRC cases using aggregated information from narratives and billing data. They postulate that rule-based systems perform well provided that features extracted using NLP are accurate. Sohn *et al*[19] built a system to extract physician-asserted side effects from EHR clinical narratives of psychiatry and psychology patients. Their system leverages NLP using cTAKES along with decision trees (C4.5) using side effect keyword features and pattern-matching rules. A comparison of rule-based and hybrid systems showed that the rule-based system had a higher F-score, with the hybrid system covering cases that were more interesting. Application of SVM to attributes with no disease-specific limitations gave a better performance than attributes clinically relevant to rheumatoid arthritis in a study by Carroll *et al.*[10] These attributes were constructed using billing codes, medication exposures, and NLP-derived concepts.

An interesting study[83] involving no custom software or rules compared performance of maximum entropy classifiers and conditional random fields, with NLP-based features extracted, from imaging and pathology reports using cTAKES, consistent with three types of cancer. Conditional random fields gave a better F-score in most cases on a physician-annotated validation set.

IBM developed the Medical Text Analysis System (MedTAS) based on their Unstructured Information Management Architecture (UIMA) framework, which leverages NLP principles and contains both rule-based and machine learning-based components. They faced challenges in implementing parts of the system using machine learning techniques because the size of their training dataset was small. Coden *et al*[71] report the pathology version of MedTAS/P tailored for extracting and storing cancer disease characteristics from pathology reports along with their relations, including temporal information and inference. Wilke *et al*[84] report a reduction in the number of false positives after augmenting results from NLP software applied to medication and laboratory data as compared with diagnostic codes alone. Kaelber *et al*[85] demonstrate the use of a commercial offering to characterize patients with venous thromboembolic events with a large deidentified dataset of 959 030 patients from 1999 to 2011.

Hybrid approaches have been successful with prediction and classification tasks in other domains. There is a growing trend in the number of studies that use such techniques. However, there is a lack of understanding regarding why a particular part of the task was performed using a particular approach. Very few studies explain the reason for performance benefits of a hybrid approach for the same task against a fully rule-based or machine learning approach.

Cohort identification systems

Initiatives such as the Electronic Medical Records and Genomics (eMERGE) network or the Cross Institutional Translational Research (CICTR) project and the Strategic Health IT Advanced Research Projects (SHARP)[86] are national level programs aimed at promoting research using EHR across multiple institutions. Such collaborative efforts are believed to be the future[87] of phenotyping of EHRs. In addition, systems such as the Stanford Translational Research Integrated Database Environment (STRIDE) at Stanford and the Duke Integrated Subject Cohort and Enrollment Research Network (DISCERN) at Duke, and others,[88 89] are specific to a single site and are instrumental in cohort identification research at these institutions. While some studies describe a general system architecture elaborating different software components of the system, certain studies address specific issues such as privacy or federated querying.

Phase I of the eMERGE network comprises Mayo Clinic, Marshfield Clinic, Northwestern University, Group Health Co-operative with the University of Washington and Vanderbilt University. They reported the complexity and heterogeneity of 14 EHR-oriented phenotyping algorithms developed as a part of the eMERGE project.[15] This study reported that the majority of the algorithms include complex, nested Boolean logic and negation, with several dependent on cardinality constraints and complex temporal logic. If whole genome scans and phenotype data extracted from EHRs are to be used across multiple institutions, then it is extremely important to standardize the representation of these data. Members of the eMERGE network describe[90] how phenotypic data from five sites studying multiple diseases were mapped to standardized terminologies and metadata repository resources, using a web-based application. Another study[55] reports the portability of an algorithm for identifying type 2 diabetes cases and controls for a genome-wide association study using data captured at the five sites and validated at three of the five sites in the network. All sites used standard metadata repositories such as ICD-9-CM for diagnostic codes, RxNorm for medications, and Logical Observation Identifiers Names and Codes (LOINC) for laboratory results.

The CICTR project was undertaken to promote collaborative research among three sites located at the University of Washington, University of California, San Francisco and University of California, Davis. Each of the member sites had independently developed their information warehouses using i2b2 as a platform. Like eMERGE, each member site used ICD-9-CM, RxNorm and LOINC for representing diagnoses, medications and laboratory results, respectively. Messages were exchanged across sites using HL7 messages. The authors[91] point out that the developed system, although functional, was considered to be over-sanitized from an end user perspective and hence lacked acceptance. This was mainly due to the aggregation, deidentification, obfuscation, and blurring of data because

of the institutional review board requirements of individual sites. The authors also comment that, although i2b2 is an extremely useful resource for setting up a data repository at an individual site, it is yet to mature for supporting multiple institutions and overcoming complex issues associated with network federation.

Studies aimed at testing portability[11 14 15 32 55 58 92–94] of phenotypic algorithms state that this is a non-trivial task[95 96] primarily because of issues with data standards.

The DISCERN project[97] at Duke University is a hybrid system that combines retrospective warehouse data and real-time clinical events exchanged using HL7 messages to alert healthcare providers of potential candidates for cohort recruitment. The typical workflow of this system consists of HL7 messages created as part of patient care, followed by processing for study-specific tasks, querying the warehouse, and finally notifying the researcher of potential candidates for the cohort study via email or text messages. The STRIDE project at Stanford University consists of a data warehouse based on HL7, an application development platform, a biospecimen data management system, and support for real-time alerting. STRIDE uses SNOMED-CT, RxNorm, ICD-9, and CPT to map important concepts and their relationships. STRIDE provides efficient access to clinical and biospecimen data for research purposes using a secure, multi-tier, service-oriented architecture. Lowe *et al*[98] developed an efficient Health Insurance Portability and Accountability Act (HIPAA)-compliant application over STRIDE, which comprises a patient cohort dashboard, clinical data-filtering capabilities, and enhanced regular expression-based searching to facilitate review of their clinical data.

TrialX[99] is a tool that matches patients to clinical trials by extracting their personal health record information from the Microsoft Health Vault and Google Health (discontinued by Google since 2011) to clinical trials using semantic web technologies. Butte *et al*[100] describe a semi-automated method for real-time recruitment of patients into a clinical trial. The system notifies a study investigator of a possible candidate for clinical trial using a simple rule. The investigator then looks into some detailed criteria for final eligibility. Embi *et al*[101] discuss a similar semi-automated clinical trial alert system and its operationalization using commercial EHR systems.

Although use of standardized terminologies is encouraged and is a well-accepted best practice, we found that not many studies adopted them. Explicit mention of these terminologies across articles is summarized in table 3.

Implications for future research

Although phenotyping has developed at a steady pace in the past few years, there is a lot of room for improvement. The findings of our study reveal many areas of research that would add value to the existing implementations. These are outlined below.

Standardization of data using terminologies such as RxNorm, SNOMED-CT, and LOINC is a critical step towards the development of phenotyping solutions portable across institutions. However, many of these terminologies have overlapping coverage of concepts, with varying degrees of granularity and sometimes even lack coverage altogether for some conditions. Articles reviewed in our study do not discuss the rationale behind the mapping of particular concepts and terminologies. While the end goal of this mapping can be decisive in choosing the right terminology, the depth of traversal in a given ontology and the need to use multiple ontologies may be topics for future research.

As discussed above, there is a lack of tools that are commonly accepted as robust and state-of-the-art and can be used in an off-the-shelf fashion for the purpose of phenotyping. In order to address this problem, the biomedical informatics community should firstly channel efforts towards making open-source tools that are well documented, maintained, and easily available to users. Secondly, there should be focus on reporting the performance of these available tools before a new one is developed.

Most articles reviewed in our study report the use of certain variables to identify patient cohorts. Often, these variables are chosen on the basis of certain clinical intuitions. However, the causality of these variables towards the phenotype is not addressed in detail. Such analysis is of great value to the clinical community. A related problem is that of developing predictive models that are explanatory in nature. Prediction techniques often use complex mathematical transformations, making it difficult to illustrate the decision-making process. This is objectionable to many physicians who are key users of systems leveraging these models. Developing visualization techniques that compliment these models is an area of research worth pursuing.

Table 3 Use of standardized terminologies

Terminology used	Number of studies
ICD-9	52
SNOMED-CT	15
RxNorm	13
CPT	13
LOINC	7

CPT, Current Procedural Terminology; ICD, International Classification of Diseases; LOINC, Logical Observation Identifiers Names and Codes; SNOMED-CT, Systematized Nomenclature of Medicine–Clinical Terms.

DISCUSSION

As this review of the literature reveals, current approaches to phenotyping are often inadequate to the task. Certainly, the use of common administrative codes such as ICD-9-CM often does not truly represent the diagnoses of a given patient population. Therefore, use of additional data sources such as those outlined in table 2 is necessary to identify these patients for many phenotyping use cases. Indeed, many studies have demonstrated that developing a model based on multiple data sources and rules can often outperform individual or isolated data-use approaches. As such, we feel there should be focus on developing systems that make holistic use of the EHR in characterizing a patient for phenotyping purposes.

The clinical NLP community has published many studies addressing complex but crucial issues such as co-reference resolution, temporal analysis, and assertion classification. However, we observed that very few studies[37 102] made use of techniques addressing these methods for the purpose of cohort identification. One possibility is the lack of widely accepted tools and techniques to perform these tasks.

Although there is a growing trend in the areas of machine learning and data mining, rule-based systems are dominant, since most analyses are pursued in individual institutions with very specific patient populations. However, if generalizable solutions are to be developed, the use of statistical and machine learning methods is necessary. Efforts to port phenotyping algorithms across multiple sites have started in recent years and should be expanded to test their robustness and scalability.

This review is not without limitations. We have presented a review of the electronic phenotyping literature from four publication venues over the period 2010–2012 and important references cited by them. While we have tried to be comprehensive, the main limitation of this review remains the vagueness of the term phenotyping, meaning that there is no exhaustive search strategy for retrieving appropriate articles across the entire available body of biomedical literature. However, we are confident that, because of the venues considered and the recent popularity of this topic, this review is representative of past and present efforts in this domain.

CONCLUSION

This review provides an overview of the different attempts at electronic phenotyping in the recent past. There are individual efforts from scratch at multiple institutions for the same task. Hence, there is a lack of well-established solutions for identifying patient cohorts. Successful projects follow certain common practices such as leveraging multiple data sources and adopting data standards. If future studies are consistent with these practices, and tools used to perform fundamental tasks are commonly accepted in the community, phenotyping using electronic medical records can progress systematically.

Contributors All authors contributed substantially towards the writing and direction of this manuscript.

Funding The project described was supported by Award Number Grant R01LM011116 from the National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Mathias JS, Gossett D, Baker DW. Use of electronic health record data to evaluate overuse of cervical cancer screening. *J Am Med Inform Assoc* 2012;19:e96–101.
2 Strom BL, Schinnar R, Jones J, *et al*. Detecting pregnancy use of non-hormonal category X medications in electronic medical records. *J Am Med Inform Assoc* 2011;18(Suppl 1):i81–6.
3 Garvin JH, DuVall SL, South BR, *et al*. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. *J Am Med Inform Assoc* 2012;19:859–66.
4 Sun J, Hu J, Luo D, *et al*. Combining knowledge and data driven insights for identifying risk factors using electronic health records. *AMIA Annu Symp Proc* 2012;2012:901–10.
5 Son C-S, Kim Y-N, Kim H-S, *et al*. Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches. *J Biomed Inform* 2012;45:999–1008.
6 Sarkar IN, Chen ES. Determining compound comorbidities for heart failure from hospital discharge data. *AMIA Annu Symp Proc* 2012;2012:809–18.
7 Pakhomov S, Weston SA, Jacobsen SJ, *et al*. Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care* 2007;13:281–8.
8 Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Rheum* 2004;51:952–7.
9 Liao KP, Cai T, Gainer V, *et al*. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)* 2010;62:1120–7.
10 Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA Annu Symp Proc* 2011;2011:189–96.
11 Carroll RJ, Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012;19:e162–9.
12 Ritchie MD, Denny JC, Crawford DC, *et al*. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;86:560–72.
13 Schildcrout JS, Basford MA, Pulley JM, *et al*. An analytical approach to characterize morbidity profile dissimilarity between distinct cohorts using electronic medical records. *J Biomed Inform* 2010;43:914–23.
14 Thompson WK, Rasmussen LV, Pacheco JA, *et al*. An evaluation of the NQF Quality Data Model for representing electronic health record driven phenotyping algorithms. *AMIA Annu Symp Proc* 2012;2012:911–20.
15 Conway M, Berg RL, Carrell D, *et al*. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. *AMIA Annu Symp Proc* 2011;2011:274–83.
16 Peissig PL, Rasmussen LV, Berg RL, *et al*. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *J Am Med Inform Assoc* 2012;19:225–34.
17 Forster AJ, Jennings A, Chow C, *et al*. A systematic review to evaluate the accuracy of electronic adverse drug event detection. *J Am Med Inform Assoc* 2012;19:31–8.
18 Koh Y, Yap CW, Li S-C. Development of a combined system for identification and classification of adverse drug reactions: Alerts Based on ADR Causality and Severity (ABACUS). *J Am Med Inform Assoc* 2010;17:720–2.
19 Sohn S, Kocher J-PA, Chute CG, *et al*. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *J Am Med Inform Assoc* 2011;18 (Suppl 1):i144–9.
20 Roden DM, Xu H, Denny JC, *et al*. Electronic medical records as a tool in clinical pharmacology: opportunities and challenges. *Clin Pharmacol Ther* 2012;91:1083–6.
21 Liu M, Wu Y, Chen Y, *et al*. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J Am Med Inform Assoc* 2012;19:e28–35.
22 Gurwitz D, Pirmohamed M. Pharmacogenomics: the importance of accurate phenotypes. *Pharmacogenomics* 2010;11:469–70.
23 Penberthy L, Brown R, Puma F, *et al*. Automated matching software for clinical trials eligibility: measuring efficiency and flexibility. *Contemp Clin Trials* 2010;31:207–17.
24 Percha B, Nassif H, Lipson J, *et al*. Automatic classification of mammography reports by BI-RADS breast tissue composition class. *J Am Med Inform Assoc* 2012;19:913–16.
25 Mani S, Chen Y, Arlinghaus LR, *et al*. Early prediction of the response of breast tumors to neoadjuvant chemotherapy using quantitative MRI and machine learning. *AMIA Annu Symp Proc* 2011;2011:868–77.
26 Tatari F, Akbarzadeh-T M-R, Sabahi A. Fuzzy-probabilistic multi agent system for breast cancer risk assessment and insurance premium assignment. *J Biomed Inform* 2012;45:1021–34.
27 Herskovic JR, Subramanian D, Cohen T, *et al*. Graph-based signal integration for high-throughput phenotyping. *BMC Bioinformatics* 2012;13(Suppl 1):S2.
28 Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. *Proc AMIA Annu Fall Symp* 1997;1997:829–33.
29 Jiang X, Kim J, Wu Y, *et al*. Selecting cases for whom additional tests can improve prognostication. *AMIA Annu Symp Proc* 2012;2012:1260–8.
30 Turchin A, Pendergrass ML, Kohane IS. DITTO—a tool for identification of patient cohorts from the text of physician notes in the electronic medical record. *AMIA Annu Symp Proc* 2005;2005:744–8.
31 Sohn S, Savova GK. Mayo clinic smoking status classification system: extensions and improvements. *AMIA Annu Symp Proc* 2009;2009:619–23.
32 Liu M, Shah A, Jiang M, *et al*. A study of transportability of an existing smoking status detection module across institutions. *AMIA Annu Symp Proc* 2012;2012:577–86.
33 Uzuner Ö, Goldstein I, Yuan Luo IK. Identifying patient smoking status from medical discharge. *J Am Med Inform Assoc* 15;2006:14–25.
34 Zeng QT, Goryachev S, Weiss S, *et al*. 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.
35 Uzuner O. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc* 2009;16:561–70.
36 Chapman WW, Fizman M, Chapman BE, *et al*. A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. *J Biomed Inform* 2001;34:4–14.
37 Bejan CA, Vanderwende L, Wurfel MM, *et al*. Assessing pneumonia identification from time-ordered narrative reports. *AMIA Annu Symp Proc* 2012;2012:1119–28.
38 Fiszman M, Chapman WW, Aronsky D, *et al*. Automatic detection of acute bacterial pneumonia from chest X-ray reports. *J Am Med Inform Assoc* 2000;7:593–604.
39 Bejan CA, Xia F, Vanderwende L, *et al*. Pneumonia identification using statistical feature selection. *J Am Med Inform Assoc* 2012;19:817–23.
40 Tu SW, Kemper CA, Lane NM, *et al*. A methodology for determining patients' eligibility for clinical trials. *Methods Inf Med* 1993;32:317–25.

41 South BR, Shen S, Chapman WW, *et al*. Analysis of false positive errors of an acute respiratory infection text classifier due to contextual features. *AMIA Summits Transl Sci Proc* 2010;2010:56–60.
42 Klompas M, Haney G, Church D, *et al*. Automated identification of acute hepatitis B using electronic medical record data to facilitate public health surveillance. *PLoS One* 2008;3:e2626.
43 DeLisle S, South B, Anthony JA, *et al*. Combining free text and structured electronic medical record entries to detect acute respiratory infections. *PLoS One* 2010;5:e13377.
44 Trick WE, Zagorski BM, Tokars JI, *et al*. Computer algorithms to detect bloodstream infections. *Emerg Infect Dis* 2004;10:1612–20.
45 Nachimuthu SK, Haug PJ. Early detection of sepsis in the emergency department using Dynamic Bayesian Networks. *AMIA Annu Symp Proc* 2012;2012:653–62.
46 Lazarus R, Klompas M, Campion FX, *et al*. Electronic support for public health: validated case finding and reporting for notifiable diseases using electronic medical data. *J Am Med Inform Assoc* 2009;16:18–24.
47 Schmiedeskamp M, Harpe S, Polk R, *et al*. Use of International Classification of Diseases, Ninth Revision, Clinical Modification codes and medication use data to identify nosocomial Clostridium difficile infection. *Infect Control Hosp Epidemiol* 2009;30:1070–6.
48 Fine AM, Reis BY, Nigrovic LE, *et al*. Use of population health data to refine diagnostic decision-making for pertussis. *J Am Med Inform Assoc* 2010;17:85–90.
49 Birman-Deych E, Waterman AD, Yan Y, *et al*. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care* 2005;43:480–5.
50 Kandula S, Zeng-treitler Q, Chen L, *et al*. A Bootstrapping Algorithm to Improve Cohort Identification using Structured Data. In: *AMIA Summit on Clinical Research Informatics*. 2011:200.
51 Perry T, Zha H, Oster ME, *et al*. Utility of a clinical support tool for outpatient evaluation of pediatric chest pain. *AMIA Annu Symp Proc* 2012;2012:726–33.
52 Berty HL, Simon M, Chapman BE. A semi-automated quantification of pulmonary artery dimensions in computed tomography angiography images. *AMIA Annu Symp Proc* 2012;2012:36–42.
53 Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *J Biomed Inform* 2011;44:859–68.
54 Nguyen AN, Lawley MJ, Hansen DP, *et al*. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *J Am Med Inform Assoc* 2010;17:440–5.
55 Kho AN, Hayes MG, Rasmussen-Torvik L, *et al*. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2012;19:212–18.
56 Hebert PL, Geiss LS, Tierney EF, *et al*. Identifying persons with diabetes using Medicare claims data. *Am J Med Qual* 1999;14:270–7.
57 Wright A, Pang J, Feblowitz JC, *et al*. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. *J Am Med Inform Assoc* 2011;18:859–67.
58 Li D, Endle CM, Murthy S, *et al*. Modeling and executing electronic health records driven phenotyping algorithms using the NQF Quality Data Model and JBoss® Drools engine. *AMIA Annu Symp Proc* 2012;2012:532–41.
59 Lee EK, Yuan F, Hirsh DA, *et al*. A clinical decision tool for predicting patient care characteristics: patients returning within 72 hours in the emergency department. *AMIA Annu Symp Proc* 2012;2012:495–504.
60 Li L, Chase HS, Patel CO, *et al*. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. *AMIA Annu Symp Proc* 2008;2008:404–8.
61 Friedman C, Shagina L, Lussier Y, *et al*. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 2004;11:392–402.
62 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *AMIA Annu Symp Proc* 2001;2001:17–21.
63 McCowan IA, Moore DC, Nguyen AN, *et al*. Collection of cancer stage data by classifying free-text medical reports. *J Am Med Inform Assoc* 2007;14:736–45.
64 Lehman L, Saeed M, Long W, *et al*. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annu Symp Proc* 2012;2012:505–11.
65 Friedlin J, Overhage M, Al-Haddad MA, *et al*. Comparing methods for identifying pancreatic cancer patients using electronic data sources. *AMIA Annu Symp Proc* 2010;2010:237–41.
66 Cui L, Bozorgi A, Lhatoo SD, *et al*. EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification. *AMIA Annu Symp Proc* 2012;2012:1191–200.
67 Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.
68 Pathak J, Kiefer RC, Chute CG. Using semantic web technologies for cohort identification from electronic health records for clinical research. *AMIA Summits Transl Sci Proc* 2012;2012:10–19.
69 Pathak J, Kiefer RC, Bielinski SJ, *et al*. Mining the Human Phenome using Semantic Web Technologies: A Case Study for Type 2 Diabetes. *AMIA Annu Symp Proc* 2012;2012:699–708.
70 Jain NL, Knirsch CA, Friedman C, *et al*. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. *Proc AMIA Annu Fall Symp* 1996;1996:542–6.
71 Coden A, Savova G, Sominsky I, *et al*. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *J Biomed Inform* 2009;42:937–49.
72 Chapman WW, Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;34:301–10.
73 Sesen MB, Kadir T, Alcantara R-B, *et al*. Survival prediction and treatment recommendation with Bayesian techniques in lung cancer. *AMIA Annu Symp Proc* 2012;2012:838–47.
74 Kawaler E, Cobian A, Peissig P, *et al*. Learning to predict post-hospitalization VTE risk from EHR data. *AMIA Annu Symp Proc* 2012;2012:436–45.
75 Van den Bulcke T, Vanden Broucke P, Van Hoof V, *et al*. Data mining methods for classification of medium-chain acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data. *J Biomed Inform* 2011;44:319–25.
76 Mani S, Chen Y, Elasy T, *et al*. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA Annu Symp Proc* 2012;2012:606–15.
77 Kim D, Shin H, Song YS, *et al*. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inform* 2012;45:1191–8.
78 Oberg R, Rasmussen L, Melski J, *et al*. Evaluation of the Google search appliance for patient cohort discovery. *AMIA Annu Symp Proc* 2008;2008:1104.
79 Wang SJ, Kalpathy-Cramer J, Kim JS, *et al*. Parametric survival models for predicting the benefit of adjuvant chemoradiotherapy in gallbladder cancer. *AMIA Annu Symp Proc* 2010;2010:847–51.
80 Kim DJ, Rockhill B, Colditz GA. Validation of the Harvard Cancer Risk Index: a prediction tool for individual cancer risk. *J Clin Epidemiol* 2004;57:332–40.
81 Denny JC, Ritchie MD, Basford MA, *et al*. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;26:1205–10.
82 Xu H, Fu Z, Shah A, *et al*. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA Annu Symp Proc* 2011;2011:1564–72.
83 D'Avolio LW, Nguyen TM, Farwell WR, *et al*. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *J Am Med Inform Assoc* 2010;17:375–82.
84 Wilke RA, Berg RL, Peissig P, *et al*. Use of an electronic medical record for the identification of research subjects with diabetes mellitus. *Clin Med Res* 2007;5:1–7.
85 Kaelber DC, Foster W, Gilder J, *et al*. Patient characteristics associated with venous thromboembolic events: a cohort study using pooled electronic health record data. *J Am Med Inform Assoc* 2012;19:965–72.
86 Rea S, Pathak J, Savova G, *et al*. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. *J Biomed Inform* 2012;45:763–71.
87 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* 2012;20:117–21.
88 Maldonado JA, Costa CM, Moner D, *et al*. Using the ResearchEHR platform to facilitate the practical application of the EHR standards. *J Biomed Inform* 2012;45:746–62.
89 Brooks CJ, Stephens JW, Price DE, *et al*. Use of a patient linked data warehouse to facilitate diabetes trial recruitment from primary care. *Prim Care Diabetes* 2009;3:245–8.
90 Pathak J, Wang J, Kashyap S, *et al*. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J Am Med Inform Assoc* 2011;18:376–86.
91 Anderson N, Abend A, Mandel A, *et al*. Implementation of a deidentified federated data network for population-based cohort discovery. *J Am Med Inform Assoc* 2012;19:e60–7.
92 Pacheco JA, Avila PC, Thompson JA, *et al*. A highly specific algorithm for identifying asthma cases and controls for genome-wide association studies. *AMIA Annu Symp Proc* 2009;2009:497–501.
93 Pathak J, Pan H, Wang J, *et al*. Evaluating Phenotypic Data Elements for Genetics and Epidemiological Research: Experiences from the eMERGE and PhenX Network Projects. *AMIA Summits Transl Sci Proc* 2011;2011:41–5.
94 Singleton KW, Hsu W, Bui AAT. Comparing predictive models of glioblastoma multiforme built using multi-institutional and local data sources. *AMIA Annu Symp Proc* 2012;2012:1385–92.
95 Wei W-Q, Leibson CL, Ransom JE, *et al*. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc* 2012;19:219–24.

96 Lim Choi Keung SN, Zhao L, Tyler E, *et al*. Cohort identification for clinical research: querying federated electronic healthcare records using controlled vocabularies and semantic types. *AMIA Summits Transl Sci Proc* 2012; 2012:9.
97 Ferranti JM, Gilbert W, McCall J, *et al*. The design and implementation of an open-source, data-driven cohort recruitment system: the Duke Integrated Subject Cohort and Enrollment Research Network (DISCERN). *J Am Med Inform Assoc* 2012;19:e68–75.
98 Lowe HJ, Weber S, Ramamoorthy N, *et al*. A Model for efficient review of clinical data by researchers within the STRIDE clinical data warehouse. In: *AMIA Summit on Clinical Research Informatics*. 2011.
99 Patel C, Gomadam K, Khan S, *et al*. TrialX: using semantic technologies to match patients to relevant clinical trials based on their personal health records. *Web Semant Sci Serv Agents World Wide Web* 2010;8:342–7.
100 Butte AJ, Weinstein DA, Kohane IS. Enrolling patients into clinical trials faster using realtime recuiting. *Proc AMIA Symp* 2000:111–15.
101 Embi PJ, Jain A, Clark J, *et al*. Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. *AMIA Annu Symp Proc* 2005;2005:231–5.
102 De Bruijn B, Cherry C, Kiritchenko S, *et al*. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc* 2011;18:557–62.