ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models

Juan M. Banda,* Martin Seneviratne,*
Tina Hernandez-Boussard, and Nigam H. Shah

Stanford Center for Biomedical Informatics Research, Stanford, California 94305, USA; email: nigam@stanford.edu

Annu. Rev. Biomed. Data Sci. 2018. 1:53–68

First published as a Review in Advance on May 23, 2018

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-080917-013315

*These authors contributed equally to this article

Keywords

electronic phenotyping, cohort building, electronic health records

Abstract

With the widespread adoption of electronic health records (EHRs), large repositories of structured and unstructured patient data are becoming available to conduct observational studies. Finding patients with specific conditions or outcomes, known as phenotyping, is one of the most fundamental research problems encountered when using these new EHR data. Phenotyping forms the basis of translational research, comparative effectiveness studies, clinical decision support, and population health analyses using routinely collected EHR data. We review the evolution of electronic phenotyping, from the early rule-based methods to the cutting edge of supervised and unsupervised machine learning models. We aim to cover the most influential papers in commensurate detail, with a focus on both methodology and implementation. Finally, future research directions are explored.

INTRODUCTION

The widespread adoption of electronic health records (EHRs) over the past decade in the United States has generated vast repositories of clinical data (1). These rich data sets enable observational research at an unprecedented scale and granularity, helping to guide clinical care, public health decision-making, and translational research (2).

One of the fundamental steps in utilizing these EHR data is identifying patients with certain characteristics of interest (either exposures or outcomes) via a process known as electronic phenotyping. The descriptions of phenotypes may be as simple as patients with type 2 diabetes or far more nuanced, such as patients with stage II prostate cancer and urinary urgency without evidence of urinary tract infection.

Identifying patients who have the characteristics of interest (i.e., patient cohort identification) is important for a diverse range of applications (**Table 1**). Phenotyping may be used for cross-sectional studies, for example, to identify the percentage of patients who receive a certain medication as a first-line antihypertensive. Specific use cases include monitoring adherence to diagnostic and treatment guidelines, such as cervical cancer screening rates (3), and epidemiological studies that guide public health interventions, such as estimating rates of *Clostridium difficile* infections (4). Phenotyping may also be used to conduct cohort and case-control analyses on routinely collected data. For example, Kaelber et al. (5) conducted a retrospective study on almost one million EHRs to determine risk factors associated with thromboembolic events, while Lependu et al. (6) demonstrated an increased risk of myocardial infarction among rheumatoid arthritis patients exposed to the anti-inflammatory medication rofecoxib. These association studies are important for research spanning pharmacovigilance (7), comparative effectiveness studies (8), and clinical risk factor analysis (9). By powering clinical decision support, these EHR-based analyses can also form the basis of a learning health system—one where clinical decisions from the level of an individual patient to the public health scale are informed by historical data (10).

Another important use case of phenotyping is translational informatics, powering the emerging transdisciplinary field of phenomics (11). Traditional genetic association studies relied on phenotype data from registries such as the Framingham Heart Study; however, it is now

Table 1 Applications of electronic phenotyping across study types

Study type	Use cases
Cross-sectional	Epidemiological research
	Hospital administration/resource allocation
	Adherence to diagnostic/treatment guidelines
	Quality measurement
Association (case-control/cohort)	Genome-wide association studies
	Pharmacovigilance
	Identifying clinical risk factors and protective factors
	Clinical decision support
	Clinical effectiveness research
	Predictive modeling
Experimental	Clinical trial recruitment
	Pragmatic trials
	Adaptive/randomized, embedded, multifactorial, adaptive platform trials

possible to link genomic data with EHR-derived phenotypes in order to conduct genome- and phenome-wide association studies (PheWAS) in a more reproducible manner, deepening our understanding of genome–phenome relationships and underpinning the vision of precision medicine (12, 13).

Finally, phenotyping can be a basis for integrating formal experimental studies into routine clinical care. Electronic phenotyping can be used to determine eligibility for clinical trial recruitment (14). The rise of adaptive trial designs, where the distribution to different treatment arms changes based on outcomes to date, and pragmatic clinical trials, where the focus is on evaluating the effectiveness of real-world interventions on less tightly curated treatment groups, can be operationalized with robust phenotyping methods that appropriately identify and assign subjects (14, 15).

Despite the multitude of use cases, identifying phenotypes in EHRs represents a significant informatics challenge because of the heterogeneity, incompleteness, and dynamic nature of EHR data (16). EHR data exist as structured data, including demographics, diagnosis codes, procedure codes, lab values, and medication exposures, and are coupled with unstructured data, including progress notes, discharge summaries, and imaging and pathology reports (2). These data are distributed irregularly over time and often fragmented across multiple institutions (17). There is significant variability between providers and sites in how data are input into EHRs—in part stemming from the multipurposing of EHRs as both clinical records and billing software (18). There is also the issue of accuracy, with previous studies showing significant variability in the accuracy of content entered by clinicians (19). Finally, due to the feedback loops created by clinical workflows, EHRs represent a dynamic system rather than a structured experimental one. The current state of an EHR affects its future state; for example, a patient's lab results affect the follow-up investigations and, in turn, the medications prescribed (20).

Therefore, the task of electronic phenotyping becomes far more challenging than a simple code search and requires sophisticated methods that can account for heterogeneity between patient records, leverage multiple data types, and ultimately fit the complex knowledge representation buried in EHR data. The task of designing portable phenotypes that can be used across different sites is an additional challenge.

In this review, we present the key methodological approaches to electronic phenotyping. We begin with the traditional approach of rule-based phenotyping, discuss text processing of clinical narratives, and then move to machine learning methods on both structured and unstructured data. We review the common frameworks that have emerged for sharing phenotype definitions and finally explore future directions in the field.

PAPER SELECTION

As recognized in previous reviews (17), the broad definition of phenotyping and the diversity of methods make it challenging to develop an all-encompassing methodology for literature review. For this work, we developed a PubMed query that found the majority of relevant papers from the most important journals and conferences between 2010 and 2017. As of September 2017, this query returned a total of 312 results, out of which we removed 199, as seen in **Figure 1** and detailed in **Table 2**. We also enhanced the results with other papers that were not found in PubMed but were detected via Google Scholar alerts, found in the references section of other papers, or recommended by word of mouth. To include the most recent work, we also added a manual curation of papers presented at both the American Medical Informatics Association's Joint Summits and its Annual Symposium. The PubMed query used is as follows:

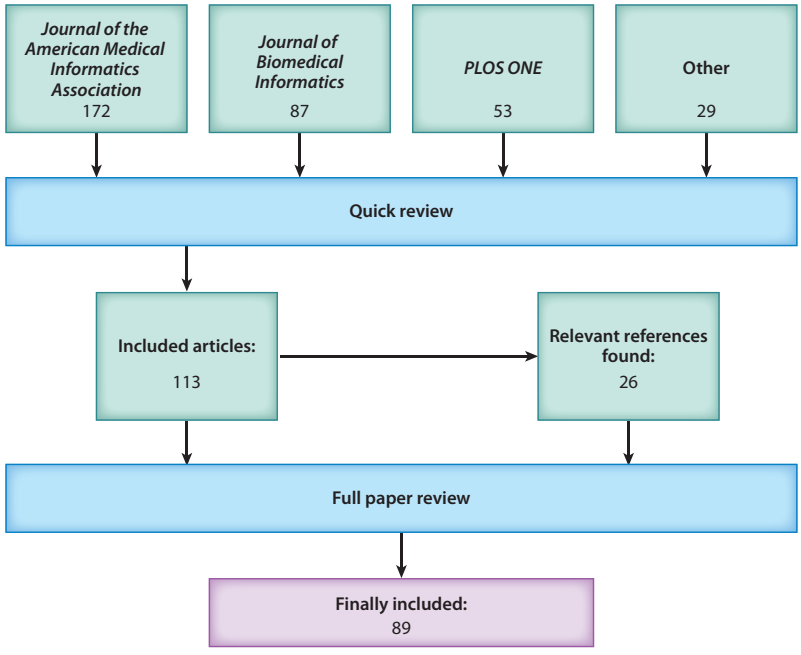

Figure 1

Paper selection process outline. Combining a PubMed query from major journals with Google Scholar alerts and other recommendations, we narrowed down the papers we reviewed in a two-step process. First, we reduced the initial number of papers (numbers inside each box in row 1) via a quick paper review. We then conducted a more thorough paper review before settling on 89 papers to discuss in this review.

```
((machine learning[MeSH Terms] OR cohort analysis[MeSH Terms] OR natural lan-
guage processing[MeSH Terms] OR phenotype[MeSH Terms]) AND electronic health
records[MeSH Terms]) AND (''2010''[Date - Publication]: ''2017''[Date - Pub-
lication])
AND (''Journal of the American Medical Informatics Association:
JAMIA''[Journal]
OR ''PloS One''[Journal]
OR ''Journal of Biomedical Informatics''[Journal])
```

Table 2 Number of papers reviewed, categorized by primary phenotyping method used

Primary method	Number of papers
Rule-based	19
Natural language processing	35
Standard machine learning	25
Learning with noisy data	11
Unsupervised phenotype discovery	11
Hybrid approaches	3
Collaborative frameworks	10
Total	89

As the medical subject heading (MeSH) terms show, we limited our search space to articles related to EHRs that were indexed with the terms "machine learning," "cohort analysis," "natural language processing," or "phenotype." We tested multiple permutations of this query using the MeSH terms as headings and subtopics and found this to give the highest yield of relevant papers.

RULE-BASED METHODS

The traditional approach to electronic phenotyping involves one or more clinicians specifying inclusion and exclusion criteria based on structured data elements such as diagnosis codes, medications, procedures, and lab values. These criteria are often drawn from consensus guidelines around diagnosis and treatment (17). A rule-based phenotype definition for finding patients with type 2 diabetes, for example, may include at least one mention of the diagnosis code, evidence of at least one hypoglycemic medication, or an HbA1c above a certain threshold (21).

Rule-based methods work well for phenotypes that have clear diagnosis and procedure codes. As early as 1999, Petersen et al. (22) could identify patients who underwent cardiac catheterization or coronary artery bypass grafting with sensitivity and specificity above 95%, using only ICD-9 (International Classification of Diseases, 9th revision) queries in a large Veterans Health Administration data set. Similarly strong results have been achieved by rule-based queries for a diverse range of conditions, including coronary artery disease (23), peripheral artery disease (24), atrial fibrillation (25), rheumatoid arthritis (26), and childhood obesity (27).

Multimodal search queries that combine diagnosis codes with other structured data fields tend to show superior performance to a single code search. Wei et al. (28) found that a single diagnosis code query was extremely variable in performance across a diverse range of disease states, from Alzheimer disease to gout to multiple sclerosis; however, using two or more codes improved average precision to 0.87. Query performance improved further when diagnosis codes were combined with medication data or keyword mentions in clinical notes—an apt demonstration of synthesizing structured and unstructured data. Schmiedeskamp et al. (29) developed queries for nosocomial *C. difficile* infections and found that adding medication data to the diagnosis codes improved positive predictive value (although with a concurrent drop in sensitivity).

The transdisciplinary field of phenomics, whose goal is to link genomic information with phenotype data extracted from EHRs, has been a major driving force for improving the quality of rule-based phenotypes (12). The eMERGE (Electronic Medical Records and Genomics) network is a consortium of collaborating academic medical centers that works to develop generalizable EHR phenotype definitions in order to conduct genome-wide association studies across shared clinical data sets (30). eMERGE is responsible for a large catalog of phenotypes, including hypothyroidism, type 2 diabetes, atrial fibrillation, and multiple sclerosis (21, 31, 32).

One of the key lessons from eMERGE is that developing phenotype criteria is an iterative process that benefits from multisite input (30), often requiring 6–8 months of development (33). The CALIBER (clinical disease research using linked bespoke studies and electronic health records) study used linked EHRs from the United Kingdom, encompassing 2.14 million patient records, to define a phenotype for atrial fibrillation (25). This required multiple cycles of clinician review and ultimately incorporated 286 codes. On a smaller scale, Kho et al. (21) developed a type 2 diabetes phenotype with a final positive predictive value greater than 98% after multiple rounds of testing and refining the algorithm across three different sites. Due to the importance of sharing and validating phenotypes in different health care settings, PheKB (Phenotype Knowledgebase; available at **http://phekb.org**) was created as a repository of phenotypes (33). There are currently 44 publicly available phenotypes on PheKB, the majority of which are rule based and use structured data; however, PheKB is now expanding into unstructured data and statistical approaches (as discussed below).

Another attempt to improve the portability of rule-based models is the ORPheUS (ontology-driven reports-based phenotyping from unique signatures) system. Yahi & Tatonetti (34) created unique signatures for 858 disease entities based solely on the high/low status of selected lab tests, with the hypothesis that these patterns may be more generalizable across sites than diagnosis codes. They demonstrated relatively good precision when validated against traditional rule-based methods; however, recall was poor (<25%).

Even with robust platforms for dissemination and testing like PheKB, the scope of rule-based approaches is limited, especially in capturing more complex phenotypes or working with less standardized data sets. Kern et al. (35) found that rule-based queries for chronic kidney disease among diabetic patients had poor sensitivity—at most 42% when using seven alternative codes. Similarly, Wei et al. (36) showed that a rule-based query for type 2 diabetes did not perform well when used at only a single site because patient data were often fragmented across inpatient and outpatient data repositories, meaning that many true positives were not identified by the prescribed phenotype rules.

Despite these limitations, rule-based methods are still in widespread use and can achieve robust results, especially if used in conjunction with unstructured data, as described in the section below.

TEXT PROCESSING FOR PHENOTYPING

Unstructured data, including text from clinical notes, discharge summaries, and radiology and pathology reports, contain a wealth of phenotypic information and represent approximately 80% of data captured in EHRs (37). The earliest methods of extracting concepts from these clinical narratives involved pattern-matching techniques against standard vocabularies (38). However, the heterogeneity of clinical expression made it difficult to reliably extract concepts (39). The last 30 years have seen the development of clinical natural language processing (NLP) techniques, which attempt to parse the semantic relationships within texts and more accurately identify phenotype attributes. Information extraction is the specific application of NLP to extract structured concepts from free text. NLP is now an integral part of the electronic phenotyping toolkit and may be used in conjunction with either rule-based methods or statistical learning methods (described below). In this section, we briefly review the broad technical categories of NLP and historical use cases, as well as key challenges and contemporary efforts to overcome them.

There are two broad families of techniques: symbolic and statistical. Symbolic techniques rely on predefined semantic relationships such as negation (40), limiting the versatility of these methods (41). Contemporary text processing is predominantly statistical, where a model is trained on an annotated corpus of text, allowing it to calibrate the importance of various semantic relationships to overall meaning. The value of such statistical text processing is that it can extract not simply a concept but also contextual cues that are important for phenotyping, such as causality and temporality (41). A thorough technical analysis of existing and emerging methods in NLP is provided elsewhere (42).

MedLEE (medical language extraction and encoding) was among the first NLP frameworks to be operationalized and integrated into the clinical information system at NewYork–Presbyterian Hospital to extract coded concepts from radiology reports (43). In the following years, a diverse range of specialized NLP pipelines have been developed for specific clinical use cases. Examples include tools for extracting respiratory diagnoses and smoking statuses from discharge summaries (44), tools for identifying peripheral artery disease in clinical notes (45) and radiology reports (46), and tools for extracting adverse outcomes following drug exposures (6). Although these techniques show excellent results in their specific use cases, the generalizability of these tools

across disciplines and across test sites is variable. Not only is there variability in clinical language between sites, but the underlying clinical data models can also vary (47). Another challenge is that such tools tend to perform best when the language of the clinical text is highly formulaic, such as in radiology and pathology reports (48). More nuanced concepts expressed in clinical notes, such as level of certainty, are more difficult to interpret and require larger volumes of accurately annotated training data.

Various attempts have been made to create interoperable NLP pipelines. The cTAKES (clinical text analysis and knowledge extraction system) is an open-source, modular NLP pipeline for clinical text, also developed at the Mayo Clinic (49). It identifies a range of clinical concepts (diagnoses, symptoms, medication exposures, etc.) with in-built functionality for detecting uncertainty and parsing temporal relationships. Inspired by the shared repositories of training text developed for nonclinical settings, Savova and coworkers (46, 48) have developed a large corpus of clinical text from the Mayo Clinic annotated with both general syntactic information such as predicate–argument relationships and medical codes from the Unified Medical Language System (UMLS). The objective is to create a large, curated training data set for research groups to create tools for different use cases.

The Open Health NLP Consortium (50) is a shared repository of NLP modules designed to be portable across sites and currently encompasses 13 tools, including multipurpose NLP frameworks such as CLAMP (Clinical Language Annotation, Modeling, and Processing Toolkit), as well as more targeted modules for extracting medication dosing (MedEx) (51), extracting cancer-specific concepts (MedKAT) (52), or parsing temporal expressions (MedTime) (53). Underpinning these advances are a deepening understanding of knowledge representations in clinical text and a move from ontologies to text-based knowledge representations (54).

In summary, although information stored in unstructured data has traditionally been underutilized in electronic phenotyping, contemporary NLP tools streamline the extraction of clinical concepts and relationships from free text. Studies developing rule-based phenotyping methods have found significant benefits in combining structured data with NLP. Carroll et al. (55) combined NLP-derived mentions of concepts and medications with structured data to create a support vector machine (SVM) classifier for rheumatoid arthritis. Liao et al. (56) demonstrated across several disease categories, from inflammatory bowel disease to multiple sclerosis, that the addition of NLP to structured data particularly improved the sensitivity of phenotypes while preserving high positive predictive value. Hence, NLP will continue to be an important part of the phenotyping toolkit, both for recognizing phenotypes directly and for processing textual content to derive features for statistical learning approaches. In the sections below, we describe the rise of statistical learning methods in phenotyping, many of which leverage both structured and unstructured data.

MACHINE LEARNING METHODS

Standard Machine Learning

In 2007, one of the first uses of machine learning for phenotyping was published by Huang et al. (57). Using a cohort of diabetic patients and controls, the authors employed a technique called FSSMC (feature selection via supervised model construction) (58). They manually distilled an initial set of 410 structured variables down to 47 features and ran FSSMC to rank the features in order of importance. They then evaluated the performance of three machine learning classifiers [naïve Bayes (59), C4.5 (60) and IB1 (61)] in identifying the diabetic patients. By exhaustive analysis, they determined the features that had the highest performance and selected those features as the most relevant ones.

As mentioned above, Carroll et al. (55) demonstrated the use of SVMs (62) to build a phenotyping model for rheumatoid arthritis. They used prescription data, ICD-9 codes, and UMLS clinical concepts extracted from clinical narratives. By testing in naïve and refined scenarios and trying different feature spaces (medications, codes, notes, or all combined), they showed that the performance of an SVM trained on their naïve data set (with all features together and no feature engineering) was almost as good as the SVM built on a refined data set. This is an important result, as it demonstrates the possibility of constructing a high-performing classifier without any feature engineering.

Using demographics, coded billing data, and lab measurements, Li et al. (63) proposed to model phenotypes with distributional association rule mining (ARM) by generating rules that are interpretable by humans and exhaustive. As ARM is a supervised machine learning algorithm, they used phenotype labels derived from their implementation of the eMERGE type 2 diabetes definition. With a label attached to each patient, they ran ARM, which uses the Apriori algorithm, to discover combinations of rules through exhaustive enumeration. By comparing with other standard machine learning algorithms (logistic regression, decision-trees, and SVM), the authors showed better performance against an expertly curated gold standard.

In 2014, Peissig et al. (64) used coded data to demonstrate that relational machine learning, more specifically, inductive logic programming, could be used successfully to identify nine different phenotypes while generating interpretable rules that are easy to read and discard. This approach improves over others that use flat feature representations and was validated against expertly curated gold standard sets and phenotype definitions from the eMERGE network.

Chen et al. (65) addressed two problems of high-throughput phenotyping: the automatic determination of phenotype topics and the portability of approaches among disparate health care systems. Using latent Dirichlet allocation, a topic model, Blei et al. (66) inferred phenotype topics from one population and then used variance analysis to evaluate the projection of the topics found in one population against another. One notable element of this approach is that by using PheWAS codes, Denny et al. (67) reduced the ICD-9 dimensional space and aimed to standardize the phenotypic topics found.

Learning with Noisy Data

The statistical phenotyping approaches described above all require manually labeled gold standard training and test data sets for model building and validation. The creation of these labeled training sets is a considerable bottleneck in the phenotyping process, as they are very expensive and time consuming to create and require domain expertise. An added issue with gold standard sets is that they are usually not portable across institutions and cannot be shared due to HIPAA (Health Insurance Portability and Accountability Act) regulations.

To alleviate the need for gold standards, both Halpern et al. (68) and Agarwal et al. (69) proposed methods that take advantage of clinical notes and other structured EHR data. The intuition underlying both approaches is that by using a large amount of imperfectly labeled training data, we can still learn good phenotype classifiers. Halpern et al. (68) introduced the notion of anchor variables, which are highly informative, clinically relevant features for a specific phenotype. The anchor variables, defined by clinical experts, may be a particular text phrase found in clinical notes, specific medication exposures, relevant laboratory tests, or diagnosis/procedure codes, and they need to satisfy certain mathematical properties. Agarwal et al. (69) relied on the presence of descriptive phrases (such as "type 2 diabetes mellitus") found in the clinical notes to assign a high-precision, albeit low-recall, phenotype label (termed a "noisy label") to a patient record. Both approaches rely on the assumption that a large volume of training data should compensate

for inaccuracies in the labels, as introduced in the domain of noise-tolerant learning (70, 71). The theory posits that by setting a bound on the labeling error and using very large amounts of training samples, models learned can be as good as those trained with small amounts of cleanly labeled (gold standard) data.

Halpern et al. (72) introduced corresponding software that handles both the feature extraction and model building, using the anchors and learn approach and the XPRESS (extraction of phenotypes from records using silver standards) framework by Agarwal et al. (73). In both approaches, the only supervised part of the process relies on selecting phenotype-specific anchors or noisy labels; all other parts of the process are automated. Both methods have been used to build phenotype models that perform as well as or better than rule-based definitions when compared against expert-generated gold standard sets. Banda et al. (74) combined the anchors and learn and XPRESS approaches into an R package called APHRODITE (automated phenotype routine for observational definition, identification, training, and evaluation), which allows both approaches to work with data sets in the Observational Medical Outcomes Partnership common data model. Banda et al. showed similar performance as the original XPRESS framework when using the same underlying data and phenotype definitions as Agarwal et al. (73).

One of the shortcomings of the anchors and learn and XPRESS approaches is that the initial set of anchors or noisy labels is identified by an expert in order for the phenotype model to be relevant. One potential improvement could be to use the approaches Yu et al. developed (75) and further refined (76) to automatically generate a list of anchors or noisy labels.

Unsupervised Phenotype Discovery

In their 2013 perspective piece, Hripcsak & Albers (20) introduced the term "high-throughput phenotyping"—a paradigm shift away from bespoke phenotypes expertly crafted from individual data sets and toward the generation of thousands of phenotypes with minimal human supervision in a scalable format. Fully embracing this new phenotyping paradigm, Ho et al. defined (77, 78) three pillars of phenotype definitions: (*a*) A phenotype represents complex interactions between several features, (*b*) the definition should be concise and understandable by a medical professional, and (*c*) the definition can be translated into new domain knowledge. The authors used a nonnegative tensor factorization technique called Limestone to generate dozens of phenotype candidates with no predefined phenotype definitions (78). These phenotype candidates are clusters of patients that exist in the data and correspond to specific medical conditions. The authors validated their approach by having a medical expert review the validity of the top automatically generated phenotype candidates.

Further refining the nonnegative tensor factorization–based approaches, Ho et al. (79) improved the Limestone technique by decomposing the observed tensor into a bias tensor, which represents the baseline characteristics found among the input patient population, and an interaction tensor, which defines the phenotypes. When applied to data from the Centers for Medicare and Medicaid Services Data Entrepreneurs' Synthetic Public Use File, this method (called Marble) yields performance improvements when the number of phenotypes is large (79).

Wang et al. (80) proposed another method, Rubrik, that incorporates medical knowledge as guidance constraints and a built-in mechanism to complete the tensors that feature missing data. This method can discover subphenotypes and scales better than previous approaches. In 2017, Henderson et al. (81) proposed further improvements over Marble with a diversified sparse nonnegative tensor factorization method, which they aptly named Granite. The goal of the improvements, which include a flexible penalized angular regularization term on the factors, was to derive phenotypes that are less overlapping (and more useful to clinicians), which would allow the method to identify rare phenotypes.

Finally, PheKnow–Cloud is a tool developed by Henderson et al. (82) that allows researchers to evaluate the output candidate phenotypes from the previously mentioned methods using medical literature. By associating the candidate phenotypes with a set of synonym terms using biomedical vocabularies [e.g., MeSH, ICD, SNOMED-CT (Systematized Nomenclature of Medicine–Clinical Terms)] and subsequently performing a co-occurrence search in PubMed Central's open access subset, the tool ranks the candidate phenotype definitions by plausibility. This approach reduces the number of candidate phenotypes that must be manually reviewed by medical experts.

HYBRID APPROACHES

Yu et al. (75) developed AFEP (automated feature extraction for phenotyping), a method that enables an automated feature selection process for high-throughput phenotyping. In AFEP, by leveraging publicly available data sources of medical knowledge like Medscape and Wikipedia, the framework produces a list of UMLS concepts that are proposed as features to use when learning a classifier for the phenotype of interest. These proposed features are further refined by only keeping the concepts found in clinical notes of EHRs (identified via NLP). Finally, ICD-9 codes and numbers of notes are added to the feature space, and an ElasticNet (83) penalized logistic regression model is built. Yu et al. (76) improved on AFEP with a framework called SAFE (surrogate-assisted feature extraction) by (*a*) using five data sources (i.e., Merck Manuals, Wikipedia, Medscape, Mayo Clinic diseases and conditions, and MedlinePlus Medical Encyclopedia) instead of two to derive candidate features, and (*b*) incorporating the idea of imperfect labeling to build intermediate phenotype models and identify and remove noninformative features before training the final phenotype classifier from a manually labeled gold standard set of patients.

FUTURE DIRECTIONS

It is clear that robust methods for electronic phenotyping are critical to leverage the rich data in EHRs for research and operational purposes (**Table 3**). Traditional rule-based definitions and manually labeled training sets still appear to be the approaches most commonly used by the phenotyping community despite their lack of portability between phenotypes and health systems. We believe that accurate phenotyping using EHRs is necessary to bring evidence to the point of care and policy makers, with high-throughput phenotyping as the centerpiece of the new portable phenotyping paradigm.

Complementing EHRs with other types of data, including registries, wearable feeds, multiomic data, imaging, and patient-reported outcomes, is a trend started by PheWAS (67). The synthesis of genomic data with EHRs has already been shown to reduce the cost of association studies by over 80% and to significantly decrease study duration (84). The same will likely be true across these multimodal data sets, allowing us to uncover new relationships by interrogating EHRs and forming the basis for more detailed phenotypes. Boland et al. introduced the term "verotype," referring to patient groups with a shared genotype, phenotype, and specific disease pattern (85). For example, Crohn disease patients with a similar pattern of disease flares and similar genetic markers might represent a single verotype. Sophisticated phenotyping techniques may allow us to discover these verotypes and then identify future patients matching a given verotype who may benefit from tailored treatment options. Future phenotyping efforts may extend these verotypes to include other omic data, behavioral patterns, etc.

Recent years have seen an increase in the popularity of machine learning approaches to phenotyping, using automated (unsupervised) or semiautomated (semisupervised) methods to enable high-throughput analyses. In this review, we have discussed some of these approaches, but there

Table 3 Selected electronic phenotyping studies and frameworks

Method/ approach	Category	EHR data used		Data sites	Number of phenotypes validated on	High-throughput	Portable
		Structured	Unstructured				
Traditional rule-based	Rule-based	X		1	Many		
eMERGE/PheKB	Rule-based	X	X	Multiple	>44		X
ORPheUS (34)	Rule-based	X		1	2		X
FSSMC (58)	Machine learning	X		1	1		
Carroll et al. (55)	Machine learning	X	X	1	1		
ARM (63)	Machine learning	X		1	1		
Peissig et al. (64)	Machine learning	X		1	9	X	
Chen et al. (65)	Machine learning	X		2	>10	X	X
Anchors and learn (72)	Machine learning	X	X	1	8	X	
XPRESS (73)	Machine learning	X	X	1	4	X	
APHRODITE (74)	Machine learning	X	X	1	2	X	X
Limestone (78)	Machine learning	X		1	NA[a]	X	
Marble (79)	Machine learning	X		1	NA[a]	X	
Rubik (80)	Machine learning	X		2	NA[a]	X	
Granite (81)	Machine learning	X		1	NA[a]	X	
AFEP (75)	Hybrid	X	X	1	2	X	
SAFE (76)	Hybrid	X	X	1	4	X	

Abbreviations: AFEP, automated feature extraction for phenotyping; APHRODITE, automated phenotype routine for observational definition, identification, training, and evaluation; ARM, association rule mining; EHR, electronic health record; eMERGE, electronic medical records and genomics; FSSMC, feature selection via supervised model construction; NA, not any; ORPheUS, ontology-driven reports-based phenotyping from unique signatures; PheKB, Phenotype Knowledgebase; SAFE, surrogate-assisted feature extraction; XPRESS, extraction of phenotypes from records using silver standards.

[a]These methods produce candidate phenotype definitions that are not usually directly validated.

are several unsolved challenges. The holy grail in phenotyping is to have a fully automated phenotyping method that produces interpretable and correct phenotype definitions that can be validated via computational mechanism. For quality purposes, it will be important to validate these unsupervised phenotypes against clinical knowledge. Traditionally, such validation has been done by expert reviewers; however, automated methods are becoming available (82). Further work is warranted around the process of validating unsupervised phenotype definitions.

We have also observed a trend toward deep learning phenotyping approaches. Previous studies have used convolutional neural networks (86), batch learning (87), and denoising auto-encoders (88, 89). There is a significant research imperative to explore representation learning to reduce the burden of feature engineering, as well as to investigate portable methods for feature engineering across phenotypes.

Collaborative research networks (90) like PCORnet (Patient-Centered Clinical Research Network) (91), i2b2 (Informatics for Integrating Biology and the Bedside) (92), MS (Mini-Sentinel) (93) and OHDSI (Observational Health Data Sciences and Informatics) (74, 94) have created pathways to share and validate phenotype models across sites. PCORnet, i2b2, MS, and OHDSI address the problem of data harmonization between multiple sites by each offering a common data

model. A shared data model is a foundational part of creating a broad phenotyping community where definitions can be tested and refined collaboratively. We expect to see the expansion of these collaborative networks, as phenotyping tools stand to benefit from shared data and diverse test sites.

In conclusion, electronic phenotyping is likely to be an active area of research in future years. We anticipate the key focus areas to be high-throughput machine learning approaches, intelligent feature engineering and dimensionality reduction, multimodal phenotype definitions based on hybrid data sets, and model-sharing communities.

DISCLOSURE STATEMENT

M.S. is an equity holder in CancerAid Pty Ltd, a patient support app for oncology care, and a former paid consultant to the Australian Digital Health Agency (2015–2016).

LITERATURE CITED

1. Pathak J, Kho AN, Denny JC. 2013. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J. Am. Med. Inform. Assoc.* 20(e2):e206–11
2. Wilcox AB. 2015. Leveraging electronic health records for phenotyping. In *Translational Informatics*, ed. PRO Payne, PJ Embi, pp. 61–74. London: Springer-Verlag
3. Mathias JS, Gossett D, Baker DW. 2012. Use of electronic health record data to evaluate overuse of cervical cancer screening. *J. Am. Med. Inform. Assoc.* 19(e1):e96–101
4. Dubberke ER, Nyazee HA, Yokoe DS, Mayer J, Stevenson KB, et al. 2012. Implementing automated surveillance for tracking *Clostridium difficile* infection at multiple healthcare facilities. *Infect. Control Hosp. Epidemiol.* 33(3):305–8
5. Kaelber DC, Foster W, Gilder J, Love TE, Jain AK. 2012. Patient characteristics associated with venous thromboembolic events: a cohort study using pooled electronic health record data. *J. Am. Med. Inform. Assoc.* 19(6):965–72
6. Lependu P, Iyer SV, Fairon C, Shah NH. 2012. Annotation analysis for testing drug safety signals using unstructured clinical notes. *J. Biomed. Semant.* 3(Suppl. 1):S5
7. Leeper NJ, Bauer-Mehren A, Iyer SV, Lependu P, Olson C, Shah NH. 2013. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. *PLOS ONE* 8(5):e63499
8. Manion FJ, Harris MR, Buyuktur AG, Clark PM, An LC, Hanauer DA. 2012. Leveraging EHR data for outcomes and comparative effectiveness research in oncology. *Curr. Oncol. Rep.* 14(6):494–501
9. Cholleti S, Post A, Gao J, Lin X, Bornstein W, et al. 2012. Leveraging derived data elements in data analytic models for understanding and predicting hospital readmissions. *Proc. AMIA Annu. Symp.* 2012:103–11
10. Longhurst CA, Harrington RA, Shah NH. 2014. A "green button" for using aggregate patient data at the point of care. *Health Aff.* 33(7):1229–35
11. Freimer N, Sabatti C. 2003. The human phenome project. *Nat. Genet.* 34(1):15–21
12. Wei W-Q, Denny JC. 2015. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med.* 7(1):41
13. Shah NH. 2013. Mining the ultimate phenome repository. *Nat. Biotechnol.* 31(12):1095–97
14. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, et al. 2013. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J. Am. Med. Inform. Assoc.* 20(e2):e226–31
15. Angus DC. 2015. Fusing randomized trials with big data: the key to self-learning health care systems? *JAMA* 314(8):767–68
16. Weiskopf NG, Weng C. 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.* 20(1):144–51
17. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, et al. 2014. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J. Am. Med. Inform. Assoc.* 21(2):221–30

18. Richesson RL, Horvath MM, Rusincovitch SA. 2014. Clinical research informatics and electronic health record data. *Yearb. Med. Inform.* 9:215–23
19. Hogan WR, Wagner MM. 1997. Accuracy of data in computer-based patient records. *J. Am. Med. Inform. Assoc.* 4(5):342–55
20. Hripcsak G, Albers DJ. 2013. Next-generation phenotyping of electronic health records. *J. Am. Med. Inform. Assoc.* 20(1):117–21
21. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, et al. 2012. 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.* 19(2):212–18
22. Petersen LA, Wright S, Normand SL, Daley J. 1999. Positive predictive value of the diagnosis of acute myocardial infarction in an administrative database. *J. Gen. Intern. Med.* 14(9):555–58
23. Esteban S, Rodríguez Tablado M, Ricci RI, Terrasa S, Kopitowski K. 2017. A rule-based electronic phenotyping algorithm for detecting clinically relevant cardiovascular disease cases. *BMC Res. Notes* 10(1):281
24. Fan J, Arruda-Olson AM, Leibson CL, Smith C, Liu G, et al. 2013. Billing code algorithms to identify cases of peripheral artery disease from administrative data. *J. Am. Med. Inform. Assoc.* 20(e2):e349–54
25. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, et al. 2014. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. *PLOS ONE* 9(11):e110900
26. Nicholson A, Ford E, Davies KA, Smith HE, Rait G, et al. 2013. Optimising use of electronic health records to describe the presentation of rheumatoid arthritis in primary care: a strategy for developing code lists. *PLOS ONE* 8(2):e54878
27. Lingren T, Thaker V, Brady C, Namjou B, Kennebeck S, et al. 2016. Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. *Appl. Clin. Inform.* 7(3):693–706
28. Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. 2016. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Inform. Assoc.* 23(e1):e20–27
29. Schmiedeskamp M, Harpe S, Polk R, Oinonen M, Pakyz A. 2009. 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.* 30(11):1070–76
30. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, et al. 2013. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* 15(10):761–71
31. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, et al. 2011. Variants near *FOXE1* are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am. J. Hum. Genet.* 89(4):529–42
32. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, et al. 2010. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* 86(4):560–72
33. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, et al. 2016. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J. Am. Med. Inform. Assoc.* 23(6):1046–52
34. Yahi A, Tatonetti NP. 2015. A knowledge-based, automated method for phenotyping in the EHR using only clinical pathology reports. *Proc. AMIA Jt. Summits Transl. Sci.* 2015:64–68
35. Kern EFO, Maney M, Miller DR, Tseng C-L, Tiwari A, et al. 2006. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. *Health Serv. Res.* 41(2):564–80
36. Wei W-Q, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, et al. 2012. 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.* 19(2):219–24
37. Martin-Sanchez F, Verspoor K. 2014. Big data in medicine is driving big changes. *Yearb. Med. Inform.* 9:14–20
38. Hersch WR, Greenes RA. 1990. SAPHIRE: an information retrieval system featuring concept-matching, automatic indexing and probabilistic retrieval. *Comput. Biomed. Res.* 23:405–20
39. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, et al. 2013. An information extraction framework for cohort identification using electronic health records. *Proc. AMIA Jt. Summits Transl. Sci.* 2013:149–53
40. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. 2001. Evaluation of negation phrases in narrative clinical reports. *Proc. AMIA Annu. Symp.* 2001:105–9

41. Nadkarni PM, Ohno-Machado L, Chapman WW. 2011. Natural language processing: an introduction. *J. Am. Med. Inform. Assoc.* 18(5):544–51
42. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, et al. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J. Biomed. Inform.* 73:14–29
43. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. 1995. Natural language processing in an operational clinical information system. *Nat. Lang. Eng.* 1(1):83–108
44. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. 2006. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med. Inform. Decis. Mak.* 6:30
45. Afzal N, Sohn S, Abram S, Scott CG, Chaudhry R, et al. 2017. Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. *J. Vasc. Surg.* 65(6):1753–61
46. Savova GK, Fan J, Ye Z, Murphy SP, Zheng J, et al. 2010. Discovering peripheral arterial disease cases from radiology notes using natural language processing. *Proc. AMIA Annu. Symp.* 2010:722–26
47. Tao C, Jiang G, Oniki TA, Freimuth RR, Zhu Q, et al. 2013. A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. *J. Am. Med. Inform. Assoc.* 20(3):554–62
48. Albright D, Lanfranchi A, Fredriksen A, Styler WF, Warner C, et al. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J. Am. Med. Inform. Assoc.* 20(5):922–30
49. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, et al. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.* 17(5):507–13
50. Masanz J, Pakhomov SV, Xu H, Wu ST, Chute CG, Liu H. 2014. Open source clinical NLP—more than any single system. *Proc. AMIA Jt. Summits Transl. Sci.* 2014:76–82
51. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. 2010. MedEx: a medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.* 17(1):19–24
52. IBM. *The MedKAT pipeline*. User Guide. **http://ohnlp.sourceforge.net/MedKATp/**
53. OHNLP (Open Health Nat. Lang. Process. Consort.). *MedTime Project Page*. User Guide, updated Nov. 18, 2013. **http://ohnlp.org/index.php/MedTime_Project_Page**
54. Deléger L, Campillos L, Ligozat A-L, Névéol A. 2017. Design of an extensive information representation scheme for clinical narratives. *J. Biomed. Semant.* 8(1):37
55. Carroll RJ, Eyler AE, Denny JC. 2011. Naïve electronic health record phenotype identification for rheumatoid arthritis. *Proc. AMIA Annu. Symp.* 2011:189–96
56. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, et al. 2015. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ* 350:h1885
57. Huang Y, McCullagh P, Black N, Harper R. 2007. Feature selection and classification model construction on type 2 diabetic patients' data. *Artif. Intell. Med.* 41(3):251–62
58. Huang Y, McCullagh PJ, Black ND. 2004. Feature selection via supervised model construction. *Proc. IEEE Int. Conf. Data Min., 4th, Brighton, U.K., 1–4 Nov.*, ed. R Rastogi, K Morik, M Bramer, X Wu, pp. 411–14. New York: IEEE
59. John GH, Langley P. 1995. Estimating continuous distributions in Bayesian classifiers. *Proc. Conf. Uncertain. Artif. Intell., 11th, Montr., Can., 18–20 Aug.*, ed. P Besnard, S Hanks, pp. 338–45. San Francisco: Morgan Kaufmann
60. Quinlan JR. 1993. *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann
61. Aha D, Kibler D. 1991. Instance-based learning algorithms. *Mach. Learn.* 6:37–66
62. Cortes C, Vapnik V. 1995. Support-vector networks. *Mach. Learn.* 20(3):273–97
63. Li D, Simon G, Chute CG, Pathak J. 2013. Using association rule mining for phenotype extraction from electronic health records. *Proc. AMIA Jt. Summ. Transl. Sci.* 2013:142–46
64. Peissig PL, Santos Costa V, Caldwell MD, Rottscheit C, Berg RL, et al. 2014. Relational machine learning for electronic health record-driven phenotyping. *J. Biomed. Inform.* 52:260–70
65. Chen Y, Ghosh J, Bejan CA, Gunter CA, Gupta S, et al. 2015. Building bridges across electronic health record systems through inferred phenotypic topics. *J. Biomed. Inform.* 55:82–93

66. Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022
67. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, et al. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* 31(12):1102–10
68. Halpern Y, Choi Y, Horng S, Sontag D. 2014. Using anchors to estimate clinical state without labeled data. *Proc. AMIA Annu. Symp.* 2014:606–15
69. Agarwal V, LePendu P, Podchiyska T, Barber R, Boland MR, et al. Using narratives as a source to automatically learn phenotype models. *Proc. Workshop Data Min. Med. Inform., 1st, Wash., D.C., 15 Nov.* **http://www.dmmh.org/dmmi2014_submission_4.pdf**
70. Simon HU. 1996. General bounds on the number of examples needed for learning probabilistic concepts. *J. Comput. Syst. Sci.* 52(2):239–54
71. Aslam JA, Decatur SE. 1996. On the sample complexity of noise-tolerant learning. *Inf. Process. Lett.* 57(4):189–95
72. Halpern Y, Horng S, Choi Y, Sontag D. 2016. Electronic medical record phenotyping using the anchor and learn framework. *J. Am. Med. Inform. Assoc.* 23(4):731–40
73. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, et al. 2016. Learning statistical models of phenotypes using noisy labeled training data. *J. Am. Med. Inform. Assoc.* 23(6):1166–73
74. Banda JM, Halpern Y, Sontag D, Shah NH. 2017. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. *Proc. AMIA Jt. Summ. Transl. Sci.* 2017:48–57
75. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, et al. 2015. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J. Am. Med. Inform. Assoc.* 22(5):993–1000
76. Yu S, Chakrabortty A, Liao KP, Cai T, Ananthakrishnan AN, et al. 2017. Surrogate-assisted feature extraction for high-throughput phenotyping. *J. Am. Med. Inform. Assoc.* 24(e1):e143–49
77. Ho JC, Ghosh J, Sun J. 2014. Extracting phenotypes from patient claim records using nonnegative tensor factorization. *Proc. Int. Conf. Brain Inform. Health, Wars., Pol., 11–14 Aug.*, ed. D Ślęzak, A-H Tan, JF Peters, L Schwabe, pp. 142–51. Cham, Switz.: Springer Int.
78. Ho JC, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, et al. 2014. Limestone: high-throughput candidate phenotype generation via tensor factorization. *J. Biomed. Inform.* 52:199–211
79. Ho JC, Ghosh J, Sun J. 2014. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. *Proc. SIGKDD Int. Conf. Knowl. Discov. Data Min., 20th, New York, N.Y., 24–27 Aug.*, pp. 115–24. New York: ACM
80. Wang Y, Chen R, Ghosh J, Denny JC, Kho A, et al. 2015. Rubik: knowledge guided tensor factorization and completion for health data analytics. *Proc. SIGKDD Int. Conf. Knowl. Discov. Data Min., 21st, Sydney, Aust., 10–13 Aug.*, pp. 1265–74. New York: ACM
81. Henderson J, Ho JC, Kho AN, Denny JC, Malin BA, et al. 2017. Granite: diversified, sparse tensor factorization for electronic health record-based phenotyping. *Proc. IEEE Int. Conf. Healthc. Inform., Park City, Utah, 23–26 Aug.*, pp. 214–23. New York: IEEE
82. Henderson J, Bridges R, Ho JC, Wallace BC, Ghosh J. 2017. PheKnow-Cloud: a tool for evaluating high-throughput phenotype candidates using online medical literature. *Proc. AMIA Jt. Summ. Transl. Sci.* 2017:149–57
83. Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67:301–20
84. Bowton E, Field JR, Wang S, Schildcrout JS, Van Driest SL, et al. 2014. Biobanks and electronic medical records: enabling cost-effective research. *Sci. Transl. Med.* 6(234):234cm3
85. Boland MR, Hripcsak G, Shen Y, Chung WK, Weng C. 2013. Defining a comprehensive verotype using electronic health records for personalized medicine. *J. Am. Med. Inform. Assoc.* 20(e2):e232–38
86. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, et al. 2017. Comparing rule-based and deep learning models for patient phenotyping. arXiv:1703.08705 [cs.CL]
87. Chiu P-H, Hripcsak G. 2017. EHR-based phenotyping: bulk learning and evaluation. *J. Biomed. Inform.* 70:35–51
88. Beaulieu-Jones BK, Greene CS. 2016. Semi-supervised learning of the electronic health record for phenotype stratification. *J. Biomed. Inform.* 64:168–78

89. Miotto R, Li L, Kidd BA, Dudley JT. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* 6:26094
90. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. 2016. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif. Intell. Med.* 71:57–61
91. Califf RM. 2014. The Patient-Centered Outcomes Research Network: a national infrastructure for comparative effectiveness research. *N.C. Med. J.* 75(3):204–10
92. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, et al. 2010. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc.* 17(2):124–30
93. McGraw D, Rosati K, Evans B. 2012. A policy framework for public health uses of electronic health data. *Pharmacoepidemiol. Drug Saf.* 21:18–22
94. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, et al. 2015. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud. Health Technol. Inform.* 216:574–78

Annual Review
of Biomedical Data
Science

Volume 1, 2018

Contents

Big Data Approaches for Modeling Response and Resistance to Cancer Drugs
Peng Jiang, William R. Sellers, and X. Shirley Liu 1

From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture
Xi Chen, Sarah A. Teichmann, and Kerstin B. Meyer 29

Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models
Juan M. Banda, Martin Seneviratne, Tina Hernandez-Boussard, and Nigam H. Shah 53

Defining Phenotypes from Clinical Data to Drive Genomic Research
Jamie R. Robinson, Wei-Qi Wei, Dan M. Roden, and Joshua C. Denny 69

Alignment-Free Sequence Analysis and Applications
Jie Ren, Xin Bai, Yang Young Lu, Kujin Tang, Ying Wang, Gesine Reinert, and Fengzhu Sun 93

Privacy Policy and Technology in Biomedical Data Science
April Moreno Arellano, Wenrui Dai, Shuang Wang, Xiaoqian Jiang, and Lucila Ohno-Machado 115

Opportunities and Challenges of Whole-Cell and -Tissue Simulations of the Outer Retina in Health and Disease
Philip J. Luthert, Luis Serrano, and Christina Kiel 131

Network Analysis as a Grand Unifier in Biomedical Data Science
Patrick McGillivray, Declan Clarke, William Meyerson, Jing Zhang, Donghoon Lee, Mengting Gu, Sushant Kumar, Holly Zhou, and Mark Gerstein 153

Deep Learning in Biomedical Data Science
Pierre Baldi 181

Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data
Pavel Sinitcyn, Jan Daniel Rudolph, and Jürgen Cox 207

Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies
Anob M. Chakrabarti, Nejc Haberman, Arne Praznik, Nicholas M. Luscombe, and Jernej Ule 235

Large-Scale Analysis of Genetic and Clinical Patient Data
Marylyn D. Ritchie 263

Visualization of Biomedical Data
Seán I. O'Donoghue, Benedetta Frida Baldi, Susan J. Clark, Aaron E. Darling, James M. Hogan, Sandeep Kaur, Lena Maier-Hein, Davis J. McCarthy, William J. Moore, Esther Stenau, Jason R. Swedlow, Jenny Vuong, and James B. Procter 275

A Census of Disease Ontologies
Melissa Haendel, Julie McMurry, Rose Relevo, Chris Mungall, Peter Robinson, and Christopher G. Chute 305

Errata

An online log of corrections to *Annual Review of Biomedical Data Science* articles may be found at http://www.annualreviews.org/errata/biodatasci