Published in final edited form as:
J Cardiovasc Transl Res. 2014 August ; 7(6): 607–614. doi:10.1007/s12265-014-9579-z.

The Electronic Health Record for Translational Research

Luke V. Rasmussen, B.S.
Northwestern University – Feinberg School of Medicine, Chicago, IL

Abstract

With growing adoption and use, the electronic health record (EHR) represents a rich source of clinical data that also offers many benefits for secondary use in biomedical research. Such benefits include access to a more comprehensive medical history, cost reductions and increased efficiency in conducting research, as well as opportunities to evaluate new and expanded populations for sufficient statistical power. Existing work utilizing EHR data has uncovered some complexities and considerations for their use, but more importantly has also generated practical lessons and solutions. Given an understanding of EHR data use in cardiovascular research, expanded adoption of this data source offers great potential to further transform the research landscape.

Keywords

Electronic health records; biomedical research; cohort identification; phenotyping; cardiovascular diseases

Introduction

The increased adoption of electronic health records (EHRs) in recent years has spurred potential benefits beyond clinical care, towards secondary uses such as for biomedical research. The National Center for Health Statistics estimates that approximately 48% of office-based physicians have a “basic” EHR system[1], with ongoing projected growth. Legislation such as the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 has been a primary motivator behind this increased adoption.

Multiple national efforts have demonstrated the utility of using EHRs for biomedical research. The electronic Medical Records and Genomics (eMERGE) network [2] is a national consortium of ten centers across the U.S. that developed and validated 14 publicly available EHR-based phenotype algorithms [3], with over 20 additional algorithms in development, to facilitate genome-wide association studies (GWAS). In addition to eMERGE, the Strategic Health IT Research Program focused on secondary use of EHRs (SHARPn) developed infrastructure to convert Quality Data Model (QDM)-based definitions into executable phenotype algorithms [4]. The Electronic Healthcare Record for

Corresponding author: Luke Rasmussen, luke.rasmussen@northwestern.edu, 312-503-2823.

Disclosures: No competing interests exist.

Human subjects/informed consent statement: No human studies were carried out by the author for this article

Animal Studies: No animal studies were carried out by the author for this article.

Clinical Research (EHR4CR) project has produced not only the technical infrastructure to integrate disparate EHRs, but also the standards and governance to ensure their interoperability, allowing for study feasibility queries [5,6]. Initiatives such as these and others have not only demonstrated the potential of EHRs as efficient and cost-effective sources of data for biomedical research [7], but have also uncovered considerations and potential challenges with this source of data, such as standardization, data quality, privacy, security and governance[8].

Benefits of EHRs for Research

Although not a panacea, the use of EHRs for research offers many benefits to researchers over (or in conjunction with) prospective data collection and paper chart reviews [9-11]. First, the EHR often represents a comprehensive medical history for a population, providing a glimpse into baselines and changes in an individual's health over time, which may not be available if data collection is purely prospective. Second, the use of the EHR offers a cost-effective option for conducting research. Since the data in an EHR is collected as part of existing medical encounters, there is no significant cost to collect data elements that already exist. Likewise, it may reduce or eliminate the need to perform procedures (i.e. laboratory tests, diagnostic imaging) to establish a baseline for a study participant. Also, EHRs allow scaling studies to large populations, which may be beneficial for rare conditions or genomic studies requiring a large population for a sufficiently powered analysis. Within cardiovascular genetic medicine, it has been previously noted that larger cohorts may be a key contribution to novel discoveries [12]. Furthermore, since the information is in an electronic format, it is amenable to computational analysis, which may speed up review and discovery. Finally, because the data collected in the EHR is broader than that needed for individual studies, the EHR as a source of information may be reused and repurposed for new hypotheses and expanded analyses.

Given the benefits demonstrated to date, the use of EHR data for cardiovascular research can be expected to increase over time. As with any data source, understanding the provenance of the data and considerations for its application to translational research will be critical for its use.

EHR Data Types

How data is stored in the EHR may vary across multiple axes, including what the data measures clinically (laboratory, medications, diagnoses, imaging, etc.), as well as the modality in which the data is persisted and retrieved for use. These factors affect how EHR data is processed by a human, as well as how they may be processed computationally. Details about the specific types of clinical information are well-described elsewhere[9], however Table 1 provides an overview of the types of data, in order of computational complexity (least to most complex).

Perhaps not surprisingly, the more complex forms of data are not as widely used as text narratives and structured data. This is in part due to the specialized knowledge needed to process those types of data, but also many times the information is abstracted into text or structured format as part of the clinical documentation.

Structured data, while arguably the easiest to compute, offers some unique challenges, primarily around the use of competing standard or local vocabularies to encode the same information, differing levels of specificity in those vocabularies, and institutional differences in how information is recorded in the EHR. This offers challenges when trying to aggregate information across multiple institutions, such as with diagnosis codes. Currently the most commonly used vocabulary in the U.S. is the International Classification of Diseases (ICD) Ninth Revision, although SNOMED CT and ICD-10 are being (or are expected to be) used. Whereas ICD-9 has fewer codes which may translate to less specificity in the meaning of a code, it is difficult to ensure semantic agreement between two seemingly equivalent codes[13]. For example, ICD-9 code 429.2 (Cardiovascular disease, unspecified) has a general equivalence code I25.10 in ICD-10 (Atherosclerotic heart disease of native coronary artery without angina pectoris). While one could say the ICD-10 code maps to the ICD-9 code (atherosclerotic heart disease is an otherwise unspecified CVD in ICD-9), the opposite is not true (not all unspecified CVD is necessarily the specific atherosclerotic heart disease represented in ICD-10). How the data is analyzed can help to address some of these differences.

Using the example codes presented, if a phenotype were being defined to exclude anyone with any history of CVD (such as with a control population), the task is easy – the more general ICD-9 code would be appropriate, as would the ICD-10 code (along with many others). If the phenotype algorithm were looking for atherosclerotic heart disease without angina pectoris, the institution using ICD-10 would be able to identify those individuals, however an institution having only ICD-9 codes may need additional supporting documentation, such as excluding all instances of the ICD-9 code that co-occur with a recorded observation of angina pectoris.

This example also illustrates that the difference in granularity of different coding systems affects not only cross-institutional implementations of an algorithm where mapping and aggregation of data is required, but also a local implementation where looking for a specific diagnosis but such specificity is not recorded.

Methods of Analysis

Many approaches may be taken to analyze EHR data [14], varying not only in complexity by the specific disease being studied, but also influenced by how associated data is stored in the EHR. As with the types of data in the EHR, analytical approaches also vary in complexity. It should be noted that no analytical approach is inherently “better” than another, and instead depends on the requirements of the study (i.e. acceptable tradeoffs in sensitivity and specificity), the capabilities of the institution, and the type of data being used from the EHR. A list of approaches is summarized in Table 2, and described in more detail below.

Rule-based solutions are perhaps the more common approach used, and consist of discrete data points from which combinations of assertions are made to arrive at a final determination (i.e. case/control status). The assertions vary in complexity, and may include nested Boolean combinations of values (has diagnosis X AND [medication Y OR lab result

Z]), temporal operators (event X at least 3 months before event Y), and arithmetic calculations on numeric values such as labs (at least an average of X over the past 5 years). The number and types of assertions needed to comprise an algorithm will vary from case to case. A selection of algorithms created by the eMERGE network, for example, had a range of 3 to 172 Boolean operators. In addition, some algorithms did not define temporal relationships, while others defined up to eight [15]. Complexity may vary by medical specialty and how the information is recorded in the EHR. For example, as diagnosis codes are recorded for billing purposes, the presence of a diagnostic code does not always guarantee the presence of a condition in a patient. To account for this, additional pieces of information can be factored in to include or exclude patients to determine their status as a case or control.

Rule-based algorithms assume that the data to be used from the EHR is in a structured format. With the advent of the Meaningful Use program, the amount of structured data points collected in the EHR can be expected to increase – both due to an increased number of EHR implementations, as well as new requirements to record structured data that is tied to standard clinical vocabularies. While there is a drive to replace clinical narrative with structured data, primarily to have more computable data, a shift to entirely structured data entry is not a well-accepted [16] or guaranteed beneficial transition [17]. Also, historical information may only be recorded as clinical narrative. Furthermore, it is not practical to record every detail of a clinical encounter in a structured format, given the complexity and richness of written text. For example, take the statement “Recent ECG is inconclusive for atrial fibrillation, despite systematic palpitations over the past 6 months”. This is a collection of a finding (palpitations, occurring over 6 months) and then a relationship between that finding and a possible diagnosis (atrial fibrillation, with a certainty of “inconclusive”). While multiple approaches exist for entering structured data, attempting to explicitly codify that phrase would require significant time and effort by a clinician. Therefore, clinical narrative text will still remain a key part of the patient's record.

This is why the use of NLP has been important to the field of EHR-based phenotyping. Natural language processing, like rule-based approaches, may vary in complexity depending on the granularity of information needed, and the manner in which it is written. The example phrase above related to atrial fibrillation represents a more complex example, where the concept “atrial fibrillation” not only needs to be extracted, but the certainty (how likely it seems the patient has the condition) as well. This also demonstrates the challenges of NLP, as the above fragment may be written by another clinician as “Evaluated ECG for afib. Palpitations persist. Results inconclusive.” requiring the computer to relate “inconclusive” back to atrial fibrillation. On the other end of the spectrum are more structured and predictable ways of recording information that follow a consistent pattern, such as “BP: 180/75”. In this case, a regular expression, which codifies a pattern, may be used. In our blood pressure example, a regular expression would essentially say “Find me text that starts with the phrase ‘BP:’, is followed by some whitespace (I don't care how much), is followed by some numbers, a forward slash, and some more numbers”.

In addition to rule-based and NLP approaches, machine learning (ML) is also a very powerful tool for phenotype development. Using statistical approaches, ML systems

evaluate large amounts of data in order to predict or classify patients given a set of variables or features. Multiple algorithms are used and are being developed, with specifics on each being outside the scope of this paper. At a higher level, two approaches may be found. One approach is “unsupervised”, in which a system is given a set of EHR data and without any further information and attempts to classify the patients. The “supervised” approach requires a human to provide the algorithm with a training set of patients along with the proper classification (i.e. is it a case or a control). Using the information from this training set, the algorithm learns the features of these classified patients, and uses them to predict the classification of patients it hasn't previously seen.

The analysis of images is also a viable approach. Although many times structured data or a clinical assessment of an image may provide the information sought, image analysis can be helpful when features not otherwise recorded need to be extracted. As previously mentioned, the types of images may significantly vary in resolution and clarity. Radiological images are one example, however handwritten notes or forms that are scanned into the EHR or captured electronically (with a pen-based tablet computer) are another [18]. The use of image analysis becomes more fruitful in EHR-based phenotypes where a large number of images may need to be processed, which precludes the use of a human reviewer, especially where the information sought is not otherwise recorded in a structured or text format.

Hybrid systems are also an option, combining one or more of the other analytical approaches. Within one review of eMERGE algorithms, 7 of 9 phenotypes used NLP in combination with rule-based logic [15]. A separate study specifically showed the importance of hybrid approaches to improve overall accuracy of algorithms, including both NLP and handwriting recognition to supplement a rule-based algorithm [19].

Use in Cardiovascular Research

Many examples exist in the biomedical literature of the varied uses of these data types and analytical approaches. In this section we explore existing work using EHR data sources for cardiovascular care, quality improvement and research, and considerations for these approaches to research-focused phenotype algorithm development overall.

Rule-based

In the United Kingdom, the cardiovascular disease (CVD) research using linked bespoke studies and electronic health records (CALIBER) project has built a large data repository aggregated from EHRs and disease registries [20]. Many of the algorithms implemented are rule-based, and although may be enriched with information from disease-specific registries, heavily utilize EHR data sources. This project noted the challenges of EHR-derived data for research, including consistency and quality issues with EHR data (missing data, incorrect data, differences in coding standards for the same data over time), both within a single EHR source and across practices. Their approach has been to transform the raw EHR data into a standardized data model for research purposes.

A single-site study using claims data to determine the presence of heart disease found reasonable accuracy, which varied based on the number of clinical events (requiring a

patient have at least two encounters decreased the prevalence by 50%) [21]. While limited to a single institution, the approach demonstrates the ability to use ICD-9 codes in combination with simple temporal constraints (having had at least X visits) to accurately detect heart disease, and also the sensitivity and specificity tradeoffs that may be made by adjusting these parameters. Later work from the same institution developed additional algorithms to identify coronary events using EHR data, showing significant improvements in accuracy when more specific diagnosis codes are available [22]. For a myocardial infarction with ST elevation (STEMI), initial algorithm iterations used a combination of ICD-9 codes, troponin levels, and structured data associated with the ECG indicating ST elevation. The algorithm was then simplified to only use more specific diagnosis codes provided by Intelligent Medical Objects (IMO). While the IMO-based algorithm required fewer data elements, it also is limited in portability to other institutions, as it requires the use of a proprietary vocabulary.

Green et al. evaluated the accuracy of body mass index (BMI) in conjunction with or in the absence of lab-based cholesterol results to calculate CVD risk [23]. They noted the lack of cholesterol results in the EHR for approximately 40% of patients, but that sufficient (although slightly elevated) risk prediction could be carried out using a lab-based score along with a requirement that the subject have been seen within the practice over the previous two years. Similarly, Dalton et al. verified the viability of CVD risk calculation using imputed missing EHR data – specifically blood pressure, total cholesterol, HDL, BMI and smoking status [24]. Both of these studies demonstrate practical approaches to dealing with the reality of missing data in the EHR.

Image Analysis

Takx et al. automated calculation of the coronary artery calcification (CAC) score using ECGgated computed tomography (CT) images to evaluate CVD risk. The automated approach had very good agreement with a manually reviewed gold standard [25]. Zhong et al. demonstrated an improved approach to measure left ventricle size using cardiac magnetic resonance imaging (MRI), with potential applications of assessing a diverse set of cardiovascular diseases [26]. Both of these works note the benefits of image analysis to automatically analyze a large corpus of images, and the benefits to researchers who may be able to leverage these approaches.

Machine Learning

Multiple studies exist that have used machine learning and statistical approaches for predictive and classification tasks, which were directly used for or can be applied to phenotype definitions. One such study used support vector machines (SVMs) to classify coronary heart disease patients, using features such as cholesterol (low-density lipoprotein, high-density lipoprotein, total and triglycerides) as well as age [27]. Another study explored the use of SVMs to predict the risk of developing CVD in the absence of blood tests, by using volume pulse [28]. Both approaches demonstrate more advanced techniques to address missing data in the EHR, which has the potential to greatly increase cohort size without sacrificing accuracy.

It is important to remember that advanced techniques (which take more time to develop and evaluate) do not always provide significant improvements. One study used ML classifiers to determine subtype of heart failure, and to predict the presence of heart failure (specifically that with preserved ejection fraction) using clinical observations [29]. This approach demonstrated favorable results and improved performance in classification tasks, but did not demonstrate significant improvements in accuracy over logistic regression. The use of ML algorithms, therefore, should be (as with any technique) used purposefully, and validated to ensure its accuracy.

Hybrid

Denny et al. conducted a GWAS related to atrioventricular conduction to assess genetic implications in PR intervals, using multiple data types to classify cases and controls [30]. The algorithm utilized natural language processing to evaluate clinical documents and ECG impressions for evidence of conditions such as heart failure and MI. The results from the NLP were coupled with ICD-9 codes, labs and medications to determine the case/control status, and achieved over 95% positive predictive value (PPV). This was the first study of its kind to utilize an EHR-based phenotype algorithm. Karnik et al. also combined coded data (diagnoses and procedures) along with the results of text searches into machine learning algorithms, identifying that the coded and text data sets performed similarly, and that there are merits of including clinical notes [31].

Turner et al. conducted a GWAS using two EHR systems to evaluate gene-gene interaction models with respect to HDL levels. The phenotype algorithm utilized HDL measurements from the EHR, and included analyses using median HDL, and a modeled HDL using statistical techniques and additional EHR-derived variables. The phenotype was primarily rule-based – including HDL lab measures, medications and diagnoses – but also included medication information derived using NLP [32]. A flowchart representation of the algorithm [33] shows the sequence of the criteria applied, including rule-based, NLP and statistical techniques, including the level of detail that is needed. In this implementation, for example, establishing the baseline HDL measure for a subject involves identifying potentially confounding diagnoses (cancer, diabetes, hyper/hypothyroidism), medication use (statins, niacin, estrogen) and events such as hospitalizations.

Considerations for Use

Three very important aspects of EHR-based phenotyping merit further discussion. The first is that all of the approaches mentioned require validation during their development. The use of a chart review (either using trained chart abstractors or a physician assessment) on a sample of the patient population is noted as critical to ensuring the accuracy of the phenotype algorithm [9,34]. The second related point is the iterative nature of EHR phenotype development. The creation of these algorithms requires expertise from multiple domains (clinical, research, informatics, data analytics) working in conjunction to propose, validate and refine the definition. Finally, through all of the effort to develop a phenotype definition, the resulting validated algorithm then represents a significant body of knowledge that other researchers may benefit from. Public repositories of these algorithms, either included with scientific publications or hosted at sites such as the Phenotype

KnowledgeBase (PheKB, http://www.phekb.org), are key to disseminating these to the scientific community.

Conclusion

Although many challenges of using EHR data have been discussed, the increased attention to EHR-based phenotyping has not only demonstrated its utility in biomedical research, but also provided a wealth of guidance to implementation, and existing algorithms that others may adopt. Within cardiovascular research, the use of EHR-derived phenotype algorithms has been shown as a rich and viable data source over disease-specific registries or study-specific prospective data collection.

Acknowledgments

Sources of Funding: Support for this work provided by NHGRI grant U01HG006388 and NCATS grant 8UL1TR000150-05.

References

1. Hsiao, CJ.; Hing, E. NCHS data brief, no 143. National Center for Health Statistics; Hyattsville, MD: 2014. Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001–2013.
2. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, Sanderson SC, Kannry J, Zinberg R, Basford MA, Brilliant M, Carey DJ, Chisholm RL, Chute CG, Connolly JJ, Crosslin D, Denny JC, Gallego CJ, Haines JL, Hakonarson H, Harley J, Jarvik GP, Kohane I, Kullo IJ, Larson EB, McCarty C, Ritchie MD, Roden DM, Smith ME, Bottinger EP, Williams MS. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genetics in medicine : official journal of the American College of Medical Genetics. 2013; 15(10):761–771.10.1038/gim.2013.72 [PubMed: 23743551]
3. Phenotype Knowledgebase (PheKB). [Accessed May 6, 2014] 2014. http://www.phekb.org/
4. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, Dligach D, Endle CM, Hart LA, Haug PJ, Huff SM, Kaggal VC, Li D, Liu H, Marchant K, Masanz J, Miller T, Oniki TA, Palmer M, Peterson KJ, Rea S, Savova GK, Stancl CR, Sohn S, Solbrig HR, Suesse DB, Tao C, Taylor DP, Westberg L, Wu S, Zhuo N, Chute CG. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. Journal of the American Medical Informatics Association : JAMIA. 2013; 20(e2):e341–348.10.1136/amiajnl-2013-001939 [PubMed: 24190931]
5. El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic PY, Jaulent MC, Daniel C. Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. Journal of biomedical informatics. 2011; 44(Suppl 1):S94–102.10.1016/j.jbi.2011.07.007 [PubMed: 21888989]
6. EHR4CR. [Accessed May 30, 2014] EHR4CR: Electronic Health Records for Clinical Research. 2014. http://www.ehr4cr.eu/
7. Bowton E, Field JR, Wang S, Schildcrout JS, Van Driest SL, Delaney JT, Cowan J, Weeke P, Mosley JD, Wells QS, Karnes JH, Shaffer C, Peterson JF, Denny JC, Roden DM, Pulley JM. Biobanks and Electronic Medical Records: Enabling Cost-Effective Research. Science Translational Medicine. 2014; 6(234):234cm233.10.1126/scitranslmed.3008604
8. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012; 13(6):395–405. [PubMed: 22549152]
9. Denny JC. Chapter 13: Mining electronic health records in the genomics era. PLoS computational biology. 2012; 8(12):e1002823.10.1371/journal.pcbi.1002823 [PubMed: 23300414]
10. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011; 12(6):417–428.10.1038/nrg2999 [PubMed: 21587298]

11. Weiner MG, Lyman JA, Murphy S, Weiner M. Electronic health records: high-quality electronic data for higher-quality clinical research. Informatics in primary care. 2007; 15(2):121–127.
12. Hershberger RE. Cardiovascular Genetic Medicine: Evolving Concepts, Rationale, and Implementation. Journal of cardiovascular translational research. 2008; 1:137–143. [PubMed: 20559908]
13. Boyd AD, Li JJ, Burton MD, Jonen M, Gardeux V, Achour I, Luo RQ, Zenku I, Bahroos N, Brown SB, Vanden Hoek T, Lussier YA. The discriminatory cost of ICD-10-CM transition between clinical specialties: metrics, case study, and mitigating tools. Journal of the American Medical Informatics Association. 2013
14. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM. A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association : JAMIA. 2014; 21(2):221–230.10.1136/amiajnl-2013-001935 [PubMed: 24201027]
15. Thompson WK, Rasmussen LV, Pacheco JA, Peissig PL, Denny JC, Kho AN, Miller A, Pathak J. An evaluation of the NQF Quality Data Model for representing Electronic Health Record driven phenotyping algorithms. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2012; 2012:911–920. [PubMed: 23304366]
16. Walsh SH. The clinician's perspective on electronic health records and how they can affect patient care. BMJ (Clinical research ed). 2004; 328(7449):1184–1187.10.1136/bmj.328.7449.1184
17. Fernando B, Kalra D, Morrison Z, Byrne E, Sheikh A. Benefits and risks of structuring and/or coding the presenting patient history in the electronic health record: systematic review. BMJ Quality & Safety. 201210.1136/bmjqs-2011-000450
18. Rasmussen LV, Peissig PL, McCarty CA, Starren J. Development of an optical character recognition pipeline for handwritten form fields from an electronic health record. Journal of the American Medical Informatics Association : JAMIA. 2012; 19(e1):e90–95.10.1136/amiajnl-2011-000182 [PubMed: 21890871]
19. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, Waudby C, Chen L, Denny JC, Wilke RA, Pathak J, Carrell D, Kho AN, Starren JB. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. Journal of the American Medical Informatics Association : JAMIA. 2012; 19(2):225–234.10.1136/amiajnl-2011-000456 [PubMed: 22319176]
20. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, Kivimaki M, Timmis AD, Smeeth L, Hemingway H. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). International journal of epidemiology. 2012; 41(6):1625–1638.10.1093/ije/dys188 [PubMed: 23220717]
21. Kottke TE, Baechler CJ, Parker ED. Accuracy of heart disease prevalence estimated from claims data compared with an electronic health record. Preventing chronic disease. 2012; 9:E141.10.5888/pcd9.120009 [PubMed: 22916996]
22. Kottke TE, Baechler CJ. An algorithm that identifies coronary and heart failure events in the electronic health record. Preventing chronic disease. 2013; 10:E29.10.5888/pcd10.120097 [PubMed: 23449283]
23. Green BB, Anderson ML, Cook AJ, Catz S, Fishman PA, McClure JB, Reid R. Using body mass index data in the electronic health record to calculate cardiovascular risk. American journal of preventive medicine. 2012; 42(4):342–347.10.1016/j.amepre.2011.12.009 [PubMed: 22424246]
24. Dalton AR, Bottle A, Soljak M, Okoro C, Majeed A, Millett C. The comparison of cardiovascular risk scores using two methods of substituting missing risk factor data in patient medical records. Informatics in primary care. 2011; 19(4):225–232.
25. Takx RAP, de Jong PA, Leiner T, Oudkerk M, de Koning HJ, Mol CP, Viergever MA, Išgum I. Automated Coronary Artery Calcification Scoring in Non-Gated Chest CT: Agreement and Reliability. PloS one. 2014; 9(3):e91239.10.1371/journal.pone.0091239 [PubMed: 24625525]
26. Zhong L, Zhang JM, Zhao X, Tan RS, Wan M. Automatic Localization of the Left Ventricle from Cardiac Cine Magnetic Resonance Imaging: A New Spectrum-Based Computer-Aided Tool. PloS one. 2014; 9(4):e92382.10.1371/journal.pone.0092382 [PubMed: 24722328]

27. Hongzong S, Tao W, Xiaojun Y, Huanxiang L, Zhide H, Mancang L, BoTao F. Support vector machines classification for discriminating coronary heart disease patients from non-coronary heart disease. The West Indian medical journal. 2007; 56(5):451–457. [PubMed: 18303759]

28. Alty SR, Millasseau SC, Chowienczyc PJ, Jakobsson A. Cardiovascular disease prediction using support vector machines.

29. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of clinical epidemiology. 2013; 66(4):398–407.10.1016/j.jclinepi.2012.11.008 [PubMed: 23384592]

30. Denny JC, Ritchie MD, Crawford DC, Schildcrout JS, Ramirez AH, Pulley JM, Basford MA, Masys DR, Haines JL, Roden DM. Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. Circulation. 2010; 122(20):2016–2021.10.1161/circulationaha.110.948828 [PubMed: 21041692]

31. Karnik S, Tan SL, Berg B, Glurich I, Zhang J, Vidaillet HJ, Page CD, Chowdhary R. Predicting atrial fibrillation and flutter using electronic health records. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference. 2012; 2012:5562–5565.10.1109/embc.2012.6347254

32. Turner SD, Berg RL, Linneman JG, Peissig PL, Crawford DC, Denny JC, Roden DM, McCarty CA, Ritchie MD, Wilke RA. Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. PloS one. 2011; 6(5):e19586.10.1371/journal.pone.0019586 [PubMed: 21589926]

33. Peissig, P.; Linneman, J. [Accessed June 30, 2014] High-Density Lipoproteins (HDL). 2012. http://phekb.org/phenotype/high-density-lipoproteins-hdl

34. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R, Pacheco JA, Rasmussen LV, Spangler L, Denny JC. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association : JAMIA. 2013; 20(e1):e147–154.10.1136/amiajnl-2012-000896 [PubMed: 23531748]

Abbreviations

CPT	Current Procedural Terminology
EHR	electronic health record
eMERGE	electronic Medical Records and Genomics
GWAS	genome-wide association study
ICD	International Classification of Diseases
NLP	natural language processing
QDM	Quality Data Model
SHARP	Strategic Health IT Research Program
SVM	support vector machine

Table 1

Modalities of EHR data

ICD-9 - International Classification of Diseases (ICD) Ninth Revision; CPT – Current Procedural Terminology

Modality	Description	Considerations for Computation	Examples
Structured/discrete data	Data that is stored in a database and is easily retrieved and computable.	Easily computable, but may require domain or institutional knowledge on proper interpretation of the data.	Atrial fibrillation (AF) diagnosis code – presence of the ICD-9 code 427.31 denotes that AF was evaluated in the patient during their encounter, but may be too broad as it may include those who were evaluated but do not truly have AF. Use of the problem list may provide a more sensitive measure; Bypass graft procedure code – given that several CPT codes exist for vascular grafts, review of the code list is needed to ascertain the correct procedure(s) of interest;
Text narrative	Clinical prose that is often written for human interpretation, or may be semi-structured (predictable location and format of information).	Varying levels of difficulty to extract information from the text. Utilizes approaches such as regular expressions and more advanced natural language processing (NLP).	CT impression – typically provides the most relevant observations from the CT scan, which may include details not captured in structured format, such as measurements (“1 cm mass”), confidence (“likely calcium buildup”), or severity (“mild blockage”); History and physical document – includes observations from the review of systems (presence of tenderness, specific heart rate measurements), onset of symptoms and family history; Pathology report – may provide more accurate measurements of a surgically removed mass;
Images	High-resolution radiological images, as well as dermatological photographs or even scanned documents. Typically stored in a specialized ancillary system, but lower resolution images may be stored directly in EHR.	Requires more advanced approaches to automatically identify and classify features in the image.	CT, MRI, x-ray – radiologic images may be re-evaluated for specific features of a secondary mass or an incidental finding that were not the primary focus of the exam; Scanned admission form – as these are stored electronically, symptoms and onset information recorded here may not be transcribed into a text note or a structured format, making the scanned document the sole source of the information;
Video	Live recording of a patient. Perhaps least prominent, but used extensively in some domains (i.e. sleep medicine).	Difficult – requires expertise in video analysis.	Overnight sleep study – video and audio recording of patient's movement, including audio cues of stopped breathing (confirmed with other measurements);

Table 2
Analytical approaches used for EHR-based phenotyping

Approach	Description	Examples
Rule-based	Relies on the presence of discrete data for the clinical variables of interest. Uses combinations of Boolean and arithmetic operators to arrive at a final phenotype definition.	Find patients with an ECG that have no prior ICD-9 code indicating heart failure (428.*).
Natural language processing	Information extraction from narrative text. This may involve extracting discrete pieces of data straight from the text, or using the context and semantics of sentences, or the document as a whole, to make a determination about a disease state.	Identify symptoms consistent with paroxysmal atrial fibrillation that are only recorded in a clinical note.
Machine learning	Utilizes algorithms, such as support vector machines (SVM) and Bayesian networks, to predict and classify phenotypes.	Given multiple clinical features (lab values, demographics, etc.) of a training set of patients, a SVM may use a statistical model to classify another set of patients as having or not having heart disease.
Image analysis	Evaluates features in clinical images to make a determination about the state of a condition or the absence/presence of a disease.	Estimate ventricle size from a radiology image.
Hybrid	Uses two or more of the other analytical approaches.	Exclude those with an ICD-9 code of heart failure, who also have no positive mention of heart failure in a clinical note.