Published in final edited form as:
Sci Transl Med. 2011 April 20; 3(79): 79re1. doi:10.1126/scitranslmed.3001807.

Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium

Abel N. Kho[1,2], **Jennifer A. Pacheco**[1], **Peggy L. Peissig**[3], **Luke Rasmussen**[3], **Katherine M. Newton**[4,5], **Noah Weston**[4], **Paul K. Crane**[6], **Jyotishman Pathak**[7], **Christopher G. Chute**[7], **Suzette J. Bielinski**[7], **Iftikhar J. Kullo**[8], **Rongling Li**[9], **Teri A. Manolio**[9], **Rex L. Chisholm**[1], and **Joshua C. Denny**[10]

[1]Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

[2]Regenstrief Institute, Inc., Indianapolis, IN 46202, USA

[3]Marshfield Clinic Research Foundation, Marshfield, WI 54449, USA

[4]Group Health Research Institute, Seattle, WA 98101, USA

[5]School of Public Health, University of Washington, Seattle, WA 98104, USA

[6]School of Medicine, University of Washington, Seattle, WA 98104, USA

[7]Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA

[8]Department of Internal Medicine, Mayo Clinic, Rochester, MN 55905, USA

[9]Office of Population Genomics, NHGRI, Bethesda, MD 20892-2152, USA

[10]Departments of Biomedical Informatics and Medicine, Vanderbilt University, Nashville, TN 37232, USA

Abstract

Clinical data in Electronic Medical Records (EMRs) is a potential source of longitudinal clinical data for research. The Electronic Medical Records and Genomics Network or eMERGE investigates whether data captured through routine clinical care using EMRs can identify disease phenotypes with sufficient positive and negative predictive values for use in genome wide association studies (GWAS). Using data from five different sets of EMRs, we have identified five disease phenotypes with positive predictive values of 73–98% and negative predictive values of 98–100%. A majority of EMRs captured key information (diagnoses, medications, laboratory tests) used to define phenotypes in a structured format. We identified natural language processing as an important tool to improve case identification rates. Efforts and incentives to increase the implementation of interoperable EMRs will markedly improve the availability of clinical data for genomics research.

To whom correspondence should be addressed: a-kho@northwestern.edu.

Author contributions: All authors participated in the design and interpretation of the experiments and results. A.K., J.A.P., P.P., L.R., K.N., N.W., P.C., J.P., C.C., S.B., R.C., J.C. participated in the acquisition and analysis of data. A.K., J.A.P., P.P., C.C., K.N., N.W., I.K., J.C., P.C., performed statistical analysis. A.K., P.P., K.N., J.C., C.C. led data collection and validation from each participating site. All authors contributed toward writing and editing the manuscript.

Competing interests: All authors declare no competing interests.

Introduction

Electronic Medical Records (EMRs) have been promoted as essential to improve healthcare quality (1–4). Although current adoption rates remain low, recent government efforts may dramatically increase the use of EMRs in clinical settings (5–9). The U.S. Centers for Medicare and Medicaid Services recently finalized a definition for “Meaningful Use” of EMRs, which defines standards for the recording and use of data in EMRs to promote quality care (10, 11). This standard, coupled with significant financial incentives and penalties, is intended to promote widespread adoption of EMRs within the U.S. healthcare system.

Understanding the strengths and limitations of current EMR data capture is crucial for identifying linkages between disease susceptibility and clinical presentation. In clinical care, EMRs serve to document clinical observations and patient-provider interactions and generate billing documentation. Clinical data captured in EMRs may have a secondary application in the research setting. In parallel with increasing EMR adoption, high throughput DNA sequencing has made available millions of DNA sequence reads for genetic investigations (12). Understanding the current feasibility of linking clinical data captured in EMRs and genome sequencing data has important implications for genetics research and the promise of personalized medicine (13–15).

Genome-wide association studies (GWAS) require accurate classification of disease phenotypes to maintain adequate statistical power (16, 17). The Electronic Medical Records and Genomics Network (eMERGE) (18) aims to determine whether data captured through routine clinical care using EMRs can identify disease phenotypes with sufficient positive and negative predictive values for application in GWAS. If successful, identification of disease phenotypes using EMR data may enable efficient and rapidly scalable genetic research. Specifically, greater efficiency may be gained by undertaking genome-wide single nucleotide polymorphism (SNP) genotyping only once. EMR data may be used to determine whether the individual associated with each DNA sample is used as a case, a control, or neither for multiple phenotypes, facilitating a GWAS for each phenotype. The marginal cost of each GWAS after the initial genotyping expense is then limited to the costs involved with establishing and validating the operational EMR-based phenotype definition and the costs of performing the association analyses. Indeed, as genotyping costs continue to rapidly decline, efficient and cost-effective means to identify phenotypic data from EMRs takes on increasing importance (19).

However, it is unclear whether current EMR implementation captures clinical data adequately to identify patients for research aimed at identifying the genetic basis of disease susceptibility. The eMERGE consortium has a unique opportunity to evaluate the utility of current EMRs for genomic research and to identify key areas for improvement. Here, we determine whether data recorded in EMRs for routine clinical care at five U.S. study sites can be used to define phenotypes for genomic research, and discuss the challenges and lessons learned in using data extracted from existing EMRs for GWAS.

Results

We analyzed EMR data collected from five eMERGE study sites to identify cases with one of five different disease phenotypes: dementia, cataracts, peripheral arterial disease, type 2 diabetes and cardiac conduction defects. Table 1 lists the primary phenotypes, biorepository description, and EMR characteristics for each study site. Three sites used an internally developed EMR system for both inpatient and outpatient care; the remaining two sites used commercial EMR systems. One site used different EMR systems for inpatient and outpatient

care. Some EMR systems captured data primarily from free text documents (unstructured data), and others from a mix of structured data collection and free text notes. Three sites used robust but different Natural Language Processing (NLP) tools to extract structured data from free text reports (20–25. Each study site had a separate DNA biorepository (linked to the EMR through a unique research identifier) to house biological samples for genotyping (26–29). With a single exception, all sites used an opt-in consent model to recruit participants into the biorepository. For our purposes, we analyzed only patients with records in both the institution's biorepository and EMR.

Data required to define the clinical gold standard for the five selected disease phenotypes across study sites most commonly required only one category of data (for example, diabetes could be defined by laboratory tests alone, and peripheral arterial disease by a single radiologic test), with one condition requiring two categories (Table 2). However, algorithms to identify the same phenotypes using EMR data required multiple categories of data, ranging from one to four categories (e.g. diagnostic information, medications, laboratory tests) with additional data categories required to identify covariates and exclusion criteria. In the example of Type 2 diabetes, the EMR derived phenotype required diagnoses, laboratory tests, and medications to identify likely Type 2 diabetes cases and used diagnoses to specifically exclude cases of Type 1 diabetes. All sites used demographic, diagnoses, and medication data in their phenotype definitions.

The three study sites using internally developed, text-based EMRs required significant natural language processing (NLP) efforts to extract concepts from free text documents, with each using a different NLP platform. At these study sites, use of NLP tools enabled disease phenotype definitions using data stored in unstructured clinical notes (e.g. ophthalmological examinations) and text-based reports (e.g. radiology test results and ECG reports). Sites without NLP tools or experience limited phenotype definitions to include only data available in a structured format data, and therefore readily extractable from the EMR

Across all five study sites, the percent of data captured and stored in a structured format consistently met or exceeded the current "Meaningful Use" Final Rule requirements (Goals for structured data capture and use defined by the Office of the National Coordinator to promote quality improvement using EMRs), with the notable exception of allergies and smoking status. Height, weight, and race/ethnicity, although satisfying the "Meaningful Use" requirements, demonstrated varying capture rates across sites. Only one institution with a vendor-based EMR had any data on family history stored in a structured format. At other sites, family history information was stored only in clinician notes and this information could not be extracted readily even with NLP. To define study phenotypes, no site required data categories with low rates of capture in the EMRs (allergies, family history).

Despite variations in categories and completeness of data capture across sites (Table 3), four out of the five study sites achieved Positive Predictive Values (PPV) of close to 100% for use of EMR data alone to identify their primary disease phenotype (Table 4). One site achieved a lower PPV of 73% using EMR data to identify cases with dementia. Absolute numbers of cases identified by EMR ranged from 747 to 2950 cases. Of sites with unselected non-cohort biorepositories, rates of case identification ranged from 3.6% to 13.4% of the total eligible population. Sites using disease-specific biorepositories had case identification rates of 26.8% to 50.3% of the total population, which, after excluding known controls, represented 71% and 90% of the cases identified through prospective cohort collection. Negative predictive values (NPV) ranged from 98–100% for the three sites generating control cases using electronic algorithms.

To assess the additional benefit of NLP, a comparison of the number of cases identified using structured data alone compared with that using both structured data and NLP was performed at one site (Vanderbilt University). At this site, the use of NLP tools identified 129% more cases of QRS duration (2950 vs. 1288) than did the use of structured data and string-matching only, while maintaining a PPV of 97%.

Discussion

In our study, data captured in EMRs for routine clinical care proved adequate to define five disease phenotypes across five different study sites with robust positive and negative predictive values. Encouragingly, several recent reports (30–32) demonstrate that GWAS based on EMR-derived phenotypes successfully replicated identification of genetic sequences associated with increased disease risk. Although we could achieve high PPVs using case identification algorithms based on data captured through routine clinical care, we note some attrition in the number of cases identified by this approach compared with disease-focused prospective case identification. In our study, electronic algorithms identified 71% and 90% of the possible cases within two prospectively collected disease cohorts. Reduction in case identification rates may be compensated by the efficiency and scalability of electronic algorithms across EMRs.

Across the five unique EMRs, diagnosis codes, medications, and laboratory tests were readily extracted to identify phenotypes for GWAS. Race/ethnicity, family history, exposure history (e.g. smoking) and environmental exposures were documented less frequently across all EMRs, and where present, often were captured in free text form (e.g. clinicians notes) and without consistent or standard nomenclature. Capturing interpreted test results that are typically not recorded as structured data elements (e.g., arterial doppler and electrocardiogram data) and clinician diagnoses (such as found on a problem list) generally required NLP. As a result, significant informatics efforts were required to tailor algorithms to each institution's EMR to accurately identify each phenotype.

Both "home-grown" and commercial EMRs demonstrated high PPV rates across the primary phenotypes. Given the far wider population using commercial EMRs in routine clinical care, this finding suggests potential for broad dissemination of our approach to identify cases and controls for genetic analyses to achieve well-powered studies, although the impact of differences among commercial EMR systems is unclear. Regardless of EMR type, study sites leveraged strengths in EMR data quality and site-specific data extraction methods to optimize phenotyping algorithms, often using data categories with a high proportion of structured data at sites without NLP capacity.

Historically, institutions with significant free text documentation in their EMRs developed or adapted robust NLP tools to extract data for further analysis (20,33,343-). NLP enabled sites to improve case finding by searching across a wider range of EMR data categories. The observation that NLP tools allowed identification of 129% more cases than were identified using purely structured data and string-matching only emphasizes the value of information captured in free text and is consistent with prior studies (35–37). As a consortium, eMERGE identified use of NLP to extract data from text documents as a critical tool to improve data quality for phenotyping. Sites with NLP experience shared best practices with other consortium sites to develop NLP capacity at all sites. However, in our study, even sites without NLP tools successfully identified their primary phenotype, and one site successfully replicated previously identified genotype-phenotype associations for five diseases, including type 2 diabetes (31). Certain phenotype identification algorithms, such as those for type 2 diabetes, were implemented without use of sophisticated NLP; other algorithms, such as those for identifying cardiac conduction problems, were implemented with a combination of

NLP and structured data extraction. This variation reflected institutional informatics capacity and a bias towards selection of phenotypes using data captured in structured formats at sites without NLP capacity. Sites without NLP capacity may be limited to identifying phenotypes using only data categories captured in structured fields. Approaches using only structured data could still achieve comparable PPVs, but would have lower case identification rates. However, efficient access to data across the entire spectrum of clinical EMRs, can compensate for lower identification rates to identify adequate numbers for genetic studies.

Some data categories consistently reflected low rates of structured data capture (Table 3). The EMRs in this study used Office of Management and Budget categories for race/ethnicity (38). In this study, low rates of documentation of race and ethnicity in the EMRs are consistent with prior studies of routine physician practice (39). However, lower rates of race and ethnicity documentation in EMRs may not significantly impact subsequent genetic studies. For genetic studies, ancestry estimates derived from genotype data are often used in primary association analyses rather than self-reported race/ethnicity, though the latter clearly adds important sociocultural information independent of genetic ancestry that may be useful in more refined analyses (40). Similarly, in our study, family history was primarily documented in clinician notes and was not readily extracted even with NLP tools. One site with a vendor based EMR featured a family history section enabling a mixture of structured and unstructured data capture, but attracted low rates of physician documentation. Our findings are consistent with prior studies, although current efforts are underway to promote standardized collection of key elements of family history within EMRs (41–43).

Environmental exposures play a significant role in expression of disease in genetically susceptible populations (44–47). Unfortunately, environmental factors, such as exposure to environmental toxins or contaminants, are rarely captured in existing EMRs, with the notable exception of smoking status. Substantial improvements in methods to collect and link environmental data to clinical data in EMRs may enable future studies of the association between disease and environment (48).

In our chart review, we identified a number of common data quality issues. Foremost, the absence of information may not reflect the absence of condition. Depending on the institution, significant care might be rendered at outside institutions and therefore would not appear in the study site's EMR. To address this limitation, we defined minimum data requirements (e.g. two documented clinical visits) to enhance the opportunity for clinical documentation beyond a single visit. We encountered instances of structured results violating acceptable ranges of possibility (e.g. a weight of 1000 kg, a height of 6 inches), requiring post-extraction censoring of impossible values. Lack of data equivalency posed challenges in merging data within a single EMR and across EMRs. Often data is imprecisely labeled such that different measures might be inappropriately mixed together. For example, laboratory tests with similar names (e.g. glucose) might represent different tests (e.g. blood glucose concentration vs urine glucose concentration). Similarly, diagnostic certainty differed depending on whether the diagnoses were entered in clinical notes or for billing purposes and differed across sites due to local billing practices (49). We identified use of data standards for EMR documentation as a necessary foundation to improve data quality and achieve data equivalence across sites. As a consortium, we used the federally endorsed Consolidated Health Informatics (CHI) standards (LOINC, ICD9/SNOMED, RxNorm) to promote data equivalency, and facilitate data sharing between sites (50–52). Phenotyping algorithms most commonly included diagnosis codes, medications, and laboratory tests, which are well covered by the CHI standards ICD9, RxNorm, and LOINC, respectively.

Our study sites represented academic medical centers or institutions with significant research programs and may have a greater focus on rigorous data collection for potential future research, limiting the generalizability of our findings to non-research oriented clinical care settings. However, recent national initiatives may promote more complete and standardized data collection across EMR-enabled clinical care settings. Greater adherence to standardized data collection may facilitate the role of EMRs in research and enable the sharing of phenotype definitions across EMR systems. The Centers for Medicare and Medicaid Services and the Office of the National Coordinator have written regulations defining "Meaningful Use" of EMRs that promote the recording of structured data and define coding standards for data categories such as diagnoses, laboratory tests, and medications. Clear documentation in EMRs is a necessary goal to achieve "Meaningful Use" and enables measurement and improvement in quality of care. Achieving this goal likewise improves the quality and volume of data available for research. Significant financial incentives for achieving meaningful use of an EMR (up to $63,750 per provider over 4 years) may increase the future availability of structured and standardized data from EMRs. Although EMR data may not capture the nuance of the human-human interaction between patient and provider, accurate and structured capture of diagnosis, laboratory test, and medication data, supplemented with text mining tools, has proved useful for identifying disease phenotypes for GWAS within the eMERGE network.

Widespread adoption of EMRs creates the potential for a quantum shift forward in the availability of longitudinal, real-world clinical data for genetics research. Our study suggests that current EMRs used for routine clinical care can be used to identify phenotypes for genetic studies. Future investment in the dissemination, standardization, and comprehensive capture of phenotypic and environmental data in EMRs will help to achieve rapidly scalable phenotyping efforts to match the proliferation of genomics data.

Methods

Each member of the eMERGE consortium selected a primary study phenotype and developed algorithms to identify the phenotype using data from their EMR. We characterized EMRs as either internally or commercially developed, and quantified the historical extent of data collection, and primary methods and tools available to define phenotypes from the EMR (Table 1). We identified the primary consent model, recruitment numbers, and demographics of each biorepository. All sites received approval from their institutional IRB for the conduct of this study.

We identified categories of EMR data used to define the five primary phenotypes (Table 2). At four of the five sites, as part of biorepository enrollment, additional data were collected on patients through an enrollment questionnaire (i.e., additional data collection outside of the clinical EMR); the fifth site (VU) used an opt-out, de-identified collection model that precluded collection of biorepository-specific information.

For each data category, we generated a measure of data completeness, defined as percent of the cohort with at least one recorded entry within the EMR for each data category. We classified the type of data in each category as either structured, unstructured (predominantly free text), or mixed. We defined structured data as numeric data or text data captured and stored in a predefined format as consistent with the current "Meaningful Use" definition. Unstructured data refer to data fields (e.g., clinical notes) that typically require subsequent processing to be useful for phenotype identification algorithms. In order to identify a comparable cohort in each EMR, we defined study patients as those enrolled within the site's biorepository who had at least two in-person visits to the healthcare institution

documented within the EMR. For the analyses presented here, study patients were not limited to those with one of the primary phenotypes.

To determine the accuracy of defining phenotypes using EMR data alone, we reviewed 100 clinical charts from the EMR at each site. Three sites used clinician chart review as the standard to confirm the primary phenotype from the records. One site used the clinical gold standard for their primary phenotype. The remaining site used trained EMR chart abstractors to confirm the primary phenotype. We measured the positive predictive value of EMR data to correctly identify cases for the primary phenotype compared with chart review (the standard). For three of the five phenotypes, we measured the negative predictive value (NPV) of EMR data to correctly identify control cases for the primary phenotype compared with the chart review standard. One of the study sites measured a quantitative trait (QRS duration, a measure of cardiac conduction) precluding measurement of an NPV. For the remaining phenotype – dementia – sufficient research quality control subjects were available from an ongoing prospective cohort study and there was concern that reliable identification of controls from EMR data would be prohibitively difficult (53, 54).

Acknowledgments

Funding: The eMERGE Network was initiated and funded by NHGRI, with additional funding from NIGMS through the following grants: U01-HG-004610 (Group Health Cooperative); U01-HG-004608 (Marshfield Clinic); U01-HG-04599 (Mayo Clinic); U01HG004609 (Northwestern University); U01-HG-04603 (Vanderbilt University, also serving as the Coordinating Center), and the State of Washington Life Sciences Discovery Fund award to the Northwest Institute of Medical Genetics. The Vanderbilt BioVU and the Synthetic Derivative were supported in part by CTSA grant 1 UL1 RR024975 from the National Center for Research Resources, National Institutes of Health.

References

1. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, Morton SC, Shekelle PG. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med. 2006; 144:742–752. [PubMed: 16702590]
2. Linder JA, Ma J, Bates DW, Middleton B, Stafford RS. Electronic health record use and the quality of ambulatory care in the United States. Arch Intern Med. 2007; 167:1400–1405. [PubMed: 17620534]
3. Walsh MN, Yancy CW, Albert NM, Curtis AB, Stough WG, Gheorghiade M, Heywood JT, McBride ML, Mehra MR, O'Connor CM, Reynolds D, Fonarow GC. Electronic health records and quality of care for heart failure. Am Heart J. 2010; 159:635–642. e631. [PubMed: 20362723]
4. Baron RJ. Quality improvement with an electronic health record: achievable, but not automatic. Ann Intern Med. 2007; 147:549–552. [PubMed: 17938393]
5. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, Shields A, Rosenbaum S, Blumenthal D. Use of Electronic Health Records in U.S. Hospitals. N Engl J Med. 2009; 360:1628–1638. [PubMed: 19321858]
6. DesRoches CM, Campbell EG, Rao SR, Donelan K, Ferris TG, Jha A, Kaushal R, Levy DE, Rosenbaum S, Shields AE, Blumenthal D. Electronic Health Records in Ambulatory Care -- A National Survey of Physicians. N Engl J Med. 2008; 359:50–60. [PubMed: 18565855]
7. U. S. Congress. American Recovery and Reinvestment Act. 2009.
8. Blumenthal D. Stimulating the Adoption of Health Information Technology. N Engl J Med. 2009; 360:1477–1479. [PubMed: 19321856]
9. Shea S, Hripcsak G. Accelerating the Use of Electronic Health Records in Physician Practices. N Engl J Med. 2010; 362:192–195. [PubMed: 20089969]
10. [accessed 8 October 2010] Meaningful Use Criteria - Final Rule. http://edocket.access.gpo.gov/2010/pdf/2010-17207.pdf/
11. Blumenthal D, Tavenner M. The Meaningful Use Regulation for Electronic Health Records. New England Journal of Medicine. 2010; 0

12. Nadler JJ, Downing GJ. Liberating Health Data for Clinical Research Applications. Science Translational Medicine. 2010; 2:18cm16.
13. Church GM. Genomes for ALL. (cover story). Scientific American. 2006; 294:47–54.
14. Burke W, Psaty BM. Personalized Medicine in the Era of Genomics. JAMA. 2007; 298:1682–1684. [PubMed: 17925520]
15. Cortese DA. A Vision of Individualized Medicine in the Context of Global Health. Clin Pharmacol Ther. 2007; 82:491–493. [PubMed: 17952101]
16. Ginsburg GS, Willard HF. Genomic and personalized medicine: foundations and applications. Translational Research. 2009; 154:277–287. [PubMed: 19931193]
17. Edwards B, Haynes C, Levenstien M, Finch S, Gordon D. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. BMC Genetics. 2005; 6:18. [PubMed: 15819990]
18. [accessed 8 October 2010] Electronic Medical Records and Genomics (eMERGE). http://www.gwas.net/
19. Murphy S, Churchill S, Bry L, Chueh H, Weiss S, Lazarus R, Zeng Q, Dubey A, Gainer V, Mendis M, Glaser J, Kohane I. Instrumenting the health care enterprise for discovery research in the genomic era. Genome Research. 2009; 19:1675–1681. [PubMed: 19602638]
20. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A General Natural-language Text Processor for Clinical Radiology. Journal of the American Medical Informatics Association. 1994; 1:161–174. [PubMed: 7719797]
21. Denny JC, Spickard A 3rd, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. J Am Med Inform Assoc. 2009; 16:806–815. [PubMed: 19717800]
22. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc. 2008; 15:25–28. [PubMed: 17947622]
23. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2010; 17:19–24. [PubMed: 20064797]
24. Wilke RA, Berg RL, Peissig P, Kitchner T, Sijercic B, McCarty CA, McCarty DJ. Use of an Electronic Medical Record for the Identification of Research Subjects with Diabetes Mellitus. CLINICAL MEDICINE & RESEARCH. 2007; 5:1–7. [PubMed: 17456828]
25. Denny JC, Miller RA, Waitman LR, Arrieta MA, Peterson JF. Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. Int J Med Inform. 2009; 78(Suppl 1):S34–42. [PubMed: 18938105]
26. McCarty CA, Peissig P, Caldwell MD, Wilke RA. The Marshfield Clinic Personalized Medicine Research Project: 2008 scientific update and lessons learned in the first 6 years. Personalized Medicine. 2008; 5:529–542.
27. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008; 84:362–369. [PubMed: 18500243]
28. [accessed 8 October 2010] The NUgene Project. http://www.nugene.org./
29. Kukull WA, Higdon R, Bowen JD, McCormick WC, Teri L, Schellenberg GD, van Belle G, Jolley L, Larson EB. Dementia and Alzheimer Disease Incidence: A Prospective Cohort Study. Arch Neurol. 2002; 59:1737–1746. [PubMed: 12433261]
30. Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. Journal of the American Medical Informatics Association. 2010; 17:568–574. [PubMed: 20819866]
31. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, Basford MA, Brown-Gentry K, Balser JR, Masys DR, Haines JL, Roden DM. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet. 2010; 86:560–572. [PubMed: 20362271]
32. 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. 122:2016–2021. [PubMed: 21041692]

33. [accessed August 2] Open Health Natural Language Processing (OHNLP) Consortium. www.ohnlp.org/
34. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, Peterson NB. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. Journal of the American Medical Informatics Association. 2010; 17:383–388. [PubMed: 20595304]
35. Friedlin J, Grannis S, Overhage JM. Using natural language processing to improve accuracy of automated notifiable disease reporting. AMIA Annu Symp Proc. 2008:207–211. [PubMed: 18999177]
36. Love TJ, Cai T, Karlson EW. Validation of Psoriatic Arthritis Diagnoses in Electronic Medical Records Using Natural Language Processing. Seminars in Arthritis and Rheumatism. 2010 In Press, Corrected Proof.
37. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care & Research. 62:1120–1127. [PubMed: 20235204]
38. [accessed 8 October 2010] Provisional Guidance on the Implementation of the 1997 Standards for Federal Data on Race and Ethnicity. http://minorityhealth.hhs.gov/templates/browse.aspx?lvl=2&lvlID=172/
39. Wynia MK, Ivey SL, Hasnain-Wynia R. Collection of data on patients' race and ethnic group by physician practices. N Engl J Med. 2010; 362:846–850. [PubMed: 20200391]
40. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]
41. Melton GB, Raman N, Chen ES, Sarkar IN, Pakhomov S, Madoff RD. Evaluation of family history information within clinical documents and adequacy of HL7 clinical statement and clinical genomics family history models for its representation: a case report. Journal of the American Medical Informatics Association. 17:337–340. [PubMed: 20442153]
42. Feero WG, Bigley MB, Brinner KM. New Standards and Enhanced Utility for Family Health History Information in the Electronic Health Record: An Update from the American Health Information Community's Family Health History Multi-Stakeholder Workgroup. Journal of the American Medical Informatics Association. 2008; 15:723–728. [PubMed: 18755994]
43. [accessed 8 October 2010] Surgeon General's Family Health History Initiative. http://www.hhs.gov/familyhistory//
44. Allen RW, Criqui MH, Diez Roux AV, Allison M, Shea S, Detrano R, Sheppard L, Wong ND, Stukovsky KH, Kaufman JD. Fine particulate matter air pollution, proximity to traffic, and aortic atherosclerosis. Epidemiology. 2009; 20:254–264. [PubMed: 19129730]
45. Diez-Roux AV. On genes, individuals, society, and epidemiology. Am J Epidemiol. 1998; 148:1027–1032. [PubMed: 9850123]
46. Diez Roux AV, Merkin SS, Arnett D, Chambless L, Massing M, Nieto FJ, Sorlie P, Szklo M, Tyroler HA, Watson RL. Neighborhood of residence and incidence of coronary heart disease. N Engl J Med. 2001; 345:99–106. [PubMed: 11450679]
47. Mujahid MS, Diez Roux AV, Morenoff JD, Raghunathan TE, Cooper RS, Ni H, Shea S. Neighborhood characteristics and hypertension. Epidemiology. 2008; 19:590–598. [PubMed: 18480733]
48. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association study (EWAS) on type 2 diabetes mellitus. PLoS One. 2010; 5:e10746. [PubMed: 20505766]
49. Schildcrout JS, Basford MA, Pulley JM, Masys DR, Roden DM, Wang D, Chute CG, Kullo IJ, Carrell D, Peissig P, Kho A, Denny JC. An analytical approach to characterize morbidity profile dissimilarity between distinct cohorts using electronic medical records. J Biomed Inform. 2010; 2010:3.
50. [accessed October 4th, 2010] Consolidated Health Informatics Initiative. http://www.hhs.gov/healthit/chiinitiative.html/

51. [accessed 8 October 2010] Logical Observations Identifers Names and Codes (LOINC). http://loinc.org//

52. [accessed 12 October 2010] RxNorm. http://www.nlm.nih.gov/research/umls/rxnorm//

53. Breitner JCS, Haneuse SJPA, Walker R, Dublin S, Crane PK, Gray SL, Larson EB. Risk of dementia and AD with prior exposure to NSAIDs in an elderly community-based cohort. Neurology. 2009; 72:1899–1905. [PubMed: 19386997]

54. Larson EB, Wang L, Bowen JD, McCormick WC, Teri L, Crane P, Kukull W. Exercise Is Associated with Reduced Risk for Incident Dementia among Persons 65 Years of Age and Older. Annals of Internal Medicine. 2006; 144:73–81. [PubMed: 16418406]

Sci Transl Med. Author manuscript; available in PMC 2013 June 23.

Table 1

Comparison of Electronic Medical Records (EMRs) and Biorepositories at five eMERGE institutions.

Institution	Biorepository Overview	Recruitment Model	Repository Size (Race/Ethnicity)	EMR Summary	Primary Phenotype	Phenotyping Methods*
Group Health (Seattle, WA)	**GHC Biobank** Alzheimer's Disease Patient Registry and Adult Changes in Thought Study	Disease specific Cohort	**4,000** (>96% Caucasian)	Comprehensive Vendor- based EMR since 2004 20+ years pharmacy data 15+ years ICD9 data	**Dementia**	Structured data extraction, Free-text searches, Manual chart review
Marshfield Clinic Research Foundation (Marshfield, WI)	**Personalized Medicine Research Project.** Geographically defined cohort within an integrated regional health care system	Population based	**20,000** (98% Caucasian)	Comprehensive internally developed EMR since 1985 75% participants have 20+ years medical history	**Cataracts**	Structured data extraction, NLP, Intelligent Character Recognition
Mayo Clinic (Rochester, MN)	**Vascular Diseases Biorepository.** Mayo Clinic Non-Invasive Vascular Laboratory & Exercise Stress Testing Lab	Disease specific cohort	**3,500** (>96% Caucasian)	Comprehensive Internally developed EMR since 1995. 40 years history of data extraction	**Peripheral Arterial Disease (PAD)**	Structured data extraction, NLP
Northwestern University (Chicago, IL)	**Nugene Project.** Northwestern affiliated hospitals and outpatient clinics	Population based	**10,000** (12% AA 8% Hispanic)	Comprehensive Vendor based inpatient (2001) and outpatient (1999)) EMRs 20+ years ICD9 data	**Type 2 diabetes**	Structured data extraction, Free-text searches
Vanderbilt University (Nashville, TN)	**BioVU:** Vanderbilt Clinic, diverse outpatient population	Population based, Opt-out consent model	**92,000** (11% AA)	Comprehensive internally developed EMR since 2000 35+ years medical history data	**Cardiac conduction**	Structured data extraction, NLP

*NLP, Natural Language Processing; Structured data extraction refers to retrieving EMR data that has been stored in a predefined format

Table 2

Data categories for defining a clinical gold standard, an EMR-derived phenotype, and covariates and exclusion criteria.

Primary Phenotype	Data Categories			Phenotyping Methods
	Clinical gold standard	**EMR-derived phenotype**	**Phenotype cohort (e.g. covariates, exclusion criteria)**	
Dementia	Demographics, Clinical notes (clinician documentation of mental status and histopathological examination data)	Diagnoses, medications	Demographics, laboratory tests, radiology reports	Structured data extraction, Free-text searches, Manual chart review
Cataracts	Clinical notes (Ophthalmologic examination)	Diagnoses, procedure codes	Demographics, medications	Structured data extraction, NLP, Intelligent Character Recognition
Peripheral Arterial Disease	Radiology test results (ankle-brachial index or arteriography)	Diagnoses, procedure codes, medications, radiology test results	Demographics	Structured data extraction, NLP
Type 2 Diabetes	Laboratory Tests	Diagnoses, laboratory tests, medications	Demographics, Laboratory tests, height, weight, family history, smoking history	Structured data extraction, Free-text searches
Cardiac Conduction	ECG measurements	ECG report results	Demographics, diagnoses, procedure codes, medications, laboratory tests	Structured data extraction, NLP

Table 3

Data completeness and type by common clinical categories.

	GHC[1]	MRCF[2]	Mayo[3]	NU[4]	VU[5]	Meaningful Use Goal
Number	2790	19771	3336	8161	81952	
Age	100% (S)	100% (S)	100% (S)	100% (S)	100% (S)	>50% (S)
Gender	100% (S)	100% (S)	100% (S)	100% (S)	100% (S)	>50% (S)
Race/Ethnicity	84% (S)	69% (S)	94% (S)	84% (S)	80% (S)	>50% (S)
Home Address	100% (M)	100% (S)	96% (S)	100% (M)	N/A*	
Height	76% (S)	96% (S)	98% (M)	92% (S)	64% (M)	>50%(S)
Weight	79% (S)	91% (S)	98% (M)	96% (S)	78% (M)	>50%(S)
Blood Pressure	59% (S)	97% (S)	62 % (M)	97% (S)	80% (M)	>50%(S)
Diagnoses	100% (S)	100% (S)	85 % (S)	99% (S)	91% (S)	>80% (S)
Laboratory Tests	100% (S)	98% (S)	81% (S)	99% (S)	97% (S)	>40% (S)
Medication	100% (S)	99% (S)	95% (M)	97% (M)	87% (M)	>80% (S)
Allergies	54% (M)	39% (M)	50% (M)	94% (M)	83% (U)	>80% (S)
Smoking History (Any)	86% (S)	77% (U)	94% (M)	73% (M)	>90%(U)	>50% (S)
Smoking History (Detailed Numeric)	54% (M)	N/A	94% (U)	12% (M)	0%	>50% (S)
Family History	20% (M)	(U)	(U)	36% (M)	(U)	

S, structured data; U, unstructured data; M, mixture of structured and unstructured data. Meaningful Use Goal for % data recorded in EMR and type listed for comparison.

*Addresses are removed from the Vanderbilt biorepository in the de-identification process.

[1]Group Health Cooperative

[2]Marshfield Clinic Research Foundation

[3]Mayo Clinic

[4]Northwestern University

[5]Vanderbilt University

Table 4

Performance of algorithms to identify cases and controls from EMRs for five primary phenotypes.

	GHC[1]	MCRF[2]	Mayo[3]	NU[4]	VU[5]
Primary Phenotype	Dementia	Cataract	Peripheral Arterial Disease	Type 2 diabetes	Cardiac conduction (Quantitative Trait)
EMR data sources to define phenotype	Diagnoses, Medications	Diagnoses, Procedures, Medications	Procedure Reports	Diagnoses, Laboratory Tests, Medications	Diagnoses, Laboratory Tests, Medications, ECG Results
Method to validate EMR phenotype	Physician Review *	Trained Chart Reviewers	Compared to Clinical Gold Standard	Physician Review	Physician Review
Number of Cases/Controls	747/2043	2642/1322	1679/1657	756/777	2950
Biospecimen #	2790	19771	3336	8161	81952
% of total biospecimen pool	26.8%	13.4%	50.3%	9.3%	3.6%
PPV (case/control)	73%	98%/98%	94%/99%	98%/100%	97%

* Review team included two physicians, a psychometrician, a neuropsychologist, and a study nurse

[1] Group Health Cooperative

[2] Marshfield Clinic Research Foundation

[3] Mayo Clinic

[4] Northwestern University

[5] Vanderbilt University