Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network

Katherine M Newton,[1] Peggy L Peissig,[2] Abel Ngo Kho,[3] Suzette J Bielinski,[4] Richard L Berg,[2] Vidhu Choudhary,[2] Melissa Basford,[5] Christopher G Chute,[6] Iftikhar J Kullo,[7] Rongling Li,[8] Jennifer A Pacheco,[3] Luke V Rasmussen,[3] Leslie Spangler,[1] Joshua C Denny[9]

[1]Group Health Research Institute, Seattle, Washington, USA
[2]Department of Biomedical Informatics, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, USA
[3]Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA
[4]Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
[5]Office of Personalized Medicine, Vanderbilt University, Nashville, Tennessee, USA
[6]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
[7]Division of Cardiovascular Diseases, Mayo Clinic, Rochester, Minnesota, USA
[8]Office of Population Genomics, National Human Genome Research Institute, Bethesda, Maryland, USA
[9]Departments of Biomedical Informatics and Medicine, Vanderbilt University, Nashville, Tennessee, USA

Correspondence to
Dr Katherine M Newton, Group Health Research Institute, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101, USA; newton.k@ghc.org

Received 12 February 2012
Accepted 5 March 2013
Published Online First 26 March 2013

To cite: Newton KM, Peissig PL, Kho AN, *et al*. *J Am Med Inform Assoc* 2013;**20**:e147–e154.

ABSTRACT

Background Genetic studies require precise phenotype definitions, but electronic medical record (EMR) phenotype data are recorded inconsistently and in a variety of formats.

Objective To present lessons learned about validation of EMR-based phenotypes from the Electronic Medical Records and Genomics (eMERGE) studies.

Materials and methods The eMERGE network created and validated 13 EMR-derived phenotype algorithms. Network sites are Group Health, Marshfield Clinic, Mayo Clinic, Northwestern University, and Vanderbilt University.

Results By validating EMR-derived phenotypes we learned that: (1) multisite validation improves phenotype algorithm accuracy; (2) targets for validation should be carefully considered and defined; (3) specifying time frames for review of variables eases validation time and improves accuracy; (4) using repeated measures requires defining the relevant time period and specifying the most meaningful value to be studied; (5) patient movement in and out of the health plan (transience) can result in incomplete or fragmented data; (6) the review scope should be defined carefully; (7) particular care is required in combining EMR and research data; (8) medication data can be assessed using claims, medications dispensed, or medications prescribed; (9) algorithm development and validation work best as an iterative process; and (10) validation by content experts or structured chart review can provide accurate results.

Conclusions Despite the diverse structure of the five EMRs of the eMERGE sites, we developed, validated, and successfully deployed 13 electronic phenotype algorithms. Validation is a worthwhile process that not only measures phenotype performance but also strengthens phenotype algorithm definitions and enhances their inter-institutional sharing.

INTRODUCTION

Electronic medical records (EMRs) hold abundant phenotype data, and government interest and promotion is driving their widespread use and adoption.[1] However, EMRs are designed to serve healthcare providers and patients by documenting patient–provider interactions and clinical observations, and generating billing documentation.[2 3] By contrast genetics research has developed predominantly within the controlled environment of research study populations with phenotypes specific to a disease domain. Thus, the EMR may be a useful tool for accelerating clinical and genetic research. Understanding the challenges of using EMR data as a source of clinical phenotypes (the presence of a specific trait, such as height or blood type, the presence of a disease, or the response to a medication) is critical to furthering the goal of repurposing EMRs for genetic research.

BACKGROUND AND SIGNIFICANCE

Genetic association studies of common clinical phenotypes require large numbers of cases and controls for adequate power,[4 5 6] and correct classification of cases (those with the trait) and controls (those without the trait) is critical for unbiased association estimates. EMR data can identify large numbers of clinical phenotypes such as disease (cases) and non-disease (controls), and quantitative traits of medical importance, with sufficient validity to power genome-wide association studies (GWAS) and other emerging types of genetic studies.[7] This has been demonstrated by the Electronic Medical Records and Genomics (eMERGE) network, created and funded by the National Human Genome Research Institute (NHGRI) to develop, disseminate, and apply approaches to combining DNA biorepositories with EMR systems for large-scale genomic studies. Successful eMERGE GWAS have included studies on red[8] and white[9] blood cell traits, atrioventricular conduction (ie, PR interval),[10] erythrocyte sedimentation rate,[11] and primary hypothyroidism[12] among others. Thus, EMRs linked to genetic data have the potential to shift the research focus from research-driven patient enrollment to high-throughput phenotyping in large patient populations, but EMRs are imperfect instruments for this use given the challenges extracting accurate phenotypes from them.[13] Phenotype validation across multiple EMR systems, preferably in different institutions, is a critical step in characterizing the types of phenotypes that the EMR can reliably provide, and establishing the utility of the EMR for GWAS. In this report we discuss lessons learned about phenotype validation during the eMERGE study and summarize the results of our validation efforts. 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 Genetic Medicine.

MATERIALS, METHODS, AND RESULTS

The five eMERGE sites are Group Health, Marshfield Clinic, Mayo Clinic, Northwestern University, and Vanderbilt University (table 1). Group Health and Marshfield Clinic are integrated care delivery systems that use commercial EMR systems, while the other three sites are fee-for-service systems that employ internally developed EMRs for inpatient and outpatient care. Detailed information about each site's data and biobank are available elsewhere.[7 14 15] Northwestern University uses one EMR system for inpatient and another for outpatient care. EMR system designs vary, but all sites' EMRs employ structured and semi-structured data, and free text (see definitions below). How specific data elements are captured varies among sites. For example, some sites have electronic pharmacy data and others collect it using natural language processing (NLP) applied to free text. At Group Health and Marshfield Clinic, additional data were collected through enrollment questionnaires and research studies, and these sites collect data from both their EMR and through billing databases when patients are seen by providers outside the healthcare system.

Data characterization

Over its first 4 years the eMERGE network selected, defined, and validated 13 phenotype algorithms (table 2). Phenotype algorithms with validation metrics are publicly available at http://www.PheKB.org. These examples were a mix of primary phenotypes identified by each site at the beginning of the study, and additional phenotypes selected by the network during the initial phase of the eMERGE study. We first identified similarities and differences of the EMR systems used at the eMERGE sites to provide an understanding of potential limitations in our ability to identify phenotypes across the five network sites.[16] We identified categories of data common to all sites (eg, age, sex, race/ethnicity, height, weight, blood pressure, inpatient/outpatient diagnosis codes, laboratory tests, medications), using the primary phenotypes to generate measures of data completeness and adequacy. To identify comparable cohorts across EMRs, we included only patients enrolled within each site's biobank with at least two recorded in-person visits. We defined data completeness as the percentage of the cohort with at least one recorded entry within each data category. This was critical to creating phenotype definitions with a reasonable likelihood of success. We further categorized the data in each category as structured (numeric or text data captured and stored in a predefined format), semi-structured (eg, section headers over free text), or free text (eg, text captured in a free form without predefined structure). We classified data as coded (structured) or not coded (free text or semi-structured text, and text found within images), using the latest definitions of Meaningful Use[1] to identify recommended national standards for EMR data capture. We also analyzed the constituent data elements of the eMERGE phenotyping algorithms, including logic use, and temporal characteristics.[17] We found that although the surface forms of these algorithms differed significantly, there was homogeneity in terms of the underlying logic used, including reliance on nested Boolean logic, temporality, and International Classification of Diseases-9-Clinical Modifications (ICD-9-CM) codes.

Phenotype selection

Each eMERGE site led the work on at least one phenotype (primary site). The network selected the first phenotypes for analysis based on the investigators' expertise and interests, the scientific importance of GWAS for the phenotype, and the feasibility of clearly identifying the phenotype within the EMR. As work progressed, additional phenotypes were suggested and considered, with the site that suggested the phenotype acting as the primary site. All sites eventually identified in their study population the presence or absence of every eMERGE phenotype.

Table 1 Comparison of electronic data available at eMERGE institutions*

	Group Health	Marshfield Clinic	Mayo Clinic	Northwestern University	Vanderbilt University
EMR vendor or development (year initiated)	**Epic (2003)**	**Internally developed: CattailsMD 1985**	**GE Centricity with internally developed modules (1995)**	**Epic (1998) Cerner (1998)**	**Internally developed: (StarChart) (1990s)**
Data availability*					
Pharmacy	1977	1992	1995	2002	1998 (inpatient)
Medications	1977	1992	1995	1998	1990
Laboratory	1988	1985	1978	2002	1995
Procedures	1977	1985	1907	1998	1990
Inpatient diagnosis codes	1972	1960	1907	2002	1990
Outpatient diagnosis codes	1984	1960	1907	1996	1990
Billing codes	1990	1985	1985	1977	
Unique features/ comments	eMERGE sample drawn from a study cohort	Structured data from EMR and insurance company is integrated into the Enterprise Data Warehouse	Vascular laboratory database is part of the EMR	Data are aggregated into an Enterprise Data Warehouse with data from Epic, Cerner, and multiple other data sources	DNA samples are linked to a de-identified version of the EMR, the Synthetic Derivative
Number of patients with genome-wide genotyping	2790	3964	3412	1932	8909

*At some study sites electronic data were available in billing and clinical databases before the adoption of an integrated electronic medical record.
eMERGE, Electronic Medical Records and Genomics; EMR, electronic medical record

Table 2 Electronic Medical Records and Genomics: validated phenotypes, participating sites, and validation approach by site

Phenotype	EMR categories to define phenotype	Challenges
Cataract	ICD-9 codes, eye exam, problem list, text, and scanned documents	Not all sites had adequate detail in EMR. Optical character recognition required for scanned records was not available at all sites
Dementia	ICD-9 codes, medications	Primary site had research-quality Alzheimer's diagnosis while others did not, compromising dementia as phenotype. Some sites had pharmacy database, others relied on NLP for pharmacy
Type 2 diabetes	ICD-9 codes, medications, laboratory tests	Difficulty handling repeated measures, differentiating type 1 from type 2 diabetes, abstracting medications from orders versus pharmacy versus NLP
Diabetic retinopathy	ICD-9 codes, laboratory tests, eye exam, problem list, text	Detailed data from eye exams not available at all sites
Resistant hypertension*	Systolic and diastolic blood pressure, medications, ICD-9 codes, free text, laboratory tests, ejection fraction	Difficulty with timing around blood pressure measures and handling repeated measures
Peripheral arterial disease	ICD-9 and CPT-4 codes, text, vascular lab criteria (ankle brachial index)	Ankle brachial index not in retrievable format in all EMRs
Primary hypothyroidism	ICD-9 and CPT-4 codes, medications, laboratory tests, text	Large number of exclusions posed challenges in developing chart review form. Person-level (lifetime) exclusion criteria were complicated by transience and time-frame limitations of the EMR (older records on paper)
Low levels of high-density lipoprotein cholesterol and baseline lipid values	Laboratory tests, medications, ICD-9 codes	Difficulty in handling repeated measures
Red blood cell indices	Laboratory tests, ICD-9 and CPT-4 codes, medications	Difficulty in handling repeated measures. Phenotype had a large number of exclusions
White blood cell indices	Laboratory tests and location of draw (eg, hospital vs clinic), ICD-9, CPT-4, and HCPCS codes, medications	Difficulty in handling repeated measures
Normal cardiac conduction (PR and QRS intervals)	Electronic ECG data, medications, NLP, ICD-9 and CPT codes, laboratory tests	Locating and mining electronic ECG data from vendor systems was difficult. Challenge asserting absence of heart disease (eg, excluding family history) or electrolyte abnormalities at the time of the ECG
Height	Height measurements, ICD-9 codes, medications, laboratory tests	Difficulty determining the normal range and handling repeated measures

All completed algorithms are available for download from http://PheKB.org.
The challenges discussed here are new observations that complement those in an earlier publication.[17]
*Genome-wide analysis not yet completed.
EMR, electronic medical record; HCPCS, health care common procedure system; NLP, natural language processing.

Validation approach

For each phenotype, the primary site developed the phenotype algorithm as a collaborative exercise between clinicians, clinical content experts, informaticians, epidemiologists, geneticists, and data experts. The clinical data within the algorithms included laboratory values, ICD-9-CM and Current Procedural Terminology (CPT)-4 codes, medications, and physical findings such as weight, height, and blood pressure. Given the differences across institutions, the developed algorithm was represented as 'pseudocode' to guide other sites in phenotype implementation, as opposed to providing source code that could be executed directly. The pseudocode was a written document that included and defined all variables needed to identify a phenotype, and the rules to combine them (ie, temporal conditions between two observations, number of observed diagnoses). The pseudocode thus provided a detailed map for data extraction. Each site then implemented the pseudocode based on their EMR structure.

Primary sites initially validated the phenotype algorithm performance (the success of the algorithm in identifying cases and controls, and meeting eligibility criteria) for their site-specific phenotypes and distributed phenotyping algorithms to other eMERGE sites for implementation. Validation reviews were accomplished via manual record review of paper or electronic records to confirm the correctness of the variables used to create the phenotype algorithm. The decision about how many cases and controls to review, and which sites would participate, was made on a case by case basis. For dementia (Group Health) and peripheral arterial disease (Mayo Clinic), a large number of cases had been confirmed for other studies. For cataract, both cases and controls had been previously reviewed at Marshfield Clinic. We supplemented these reviews with reviews at other sites. Because of an institutional interest in algorithm validation, Marshfield Clinic participated in validation for almost every algorithm. For the other algorithms, sites volunteered to review 50–200 subjects (persons enrolled at the institution and classified by the algorithm as having or not having the trait). The number reviewed was determined based on the collective perception of the complexities of the algorithm—a greater number of reviews was done for more complex algorithms—but in truth, this decision was somewhat arbitrary and evolved with the investigator's experience in algorithm validation.

Lessons learned from phenotype algorithm development and validation

Variable selection and definition

Selection of variables for validation

For some phenotypes (cataract, dementia, type 2 diabetes, peripheral arterial disease, hypothyroidism, resistant hypertension, and diabetic retinopathy) the goal of validation was to confirm the accuracy of case and control status. Thus, our targets for validation were the characteristics of, and inclusion and exclusion criteria for, cases and controls. For other phenotypes (QRS, low low-density lipoprotein, white blood cell count, red blood cell (RBC) count, height, lipids) the goal of the GWAS was to identify differences within normal ranges of values; controls were unnecessary, and the goal of validation was to ensure that the algorithm appropriately included those who were eligible.

We found that some phenotype algorithms were more inherently prone to error than others. Algorithms for phenotypes such as type 2 diabetes and resistant hypertension, which included a large number of variables (ICD-9-CM codes, laboratory measures, medications), required validation by review of clinical charts to understand the final determination of the diagnosing physician. In contrast, phenotypes for quantitative traits such as blood pressure and laboratory measurements were accepted as recorded in the EMR without review, except for a focused review of outliers. While extraction of quantitative traits was straightforward, validation of the patient population from whom the values were drawn could be difficult, and the validation focused on ascertaining that the patient population did not have any of the exclusion diagnoses, which could be numerous. Decisions about outliers can sometimes be made without medical record review (eg, 'serum' potassium of 50 mEq/l is incompatible with life and likely represents urine potassium). We learned that each element in the phenotype definition needed to be reviewed to determine which should be included in the validation and which could be accepted as accurately recorded in the EMR. For example, for the Mayo Clinic phenotype of RBC indices we developed an algorithm to identify trait values that could be affected by comorbidity, trauma, or drugs.[8] The algorithm was based on ICD-9 codes for hematologic disorders, solid organ malignancies, bone marrow/solid organ transplantation, hereditary anemia, and major surgery or recent trauma, as well as an NLP definition for relevant medication use.

Time periods for review of validated items

The time periods for available data varied across sites. For example, Group Health had an elderly cohort taken from a study on aging (age 65 at study entry) and electronic medication data since 1977, while Northwestern had a younger population (mean age 52) and medication data since 1996. Furthermore, institutions and departments within institutions differed in when they began using EMRs, and many patients had both paper records and EMRs. Events that occur before EMR implementation may be missed if no associated electronic data element exists. For example, at the Mayo Clinic, thyroidectomies that predated the EMR were identified during validation of the hypothyroidism phenotype. More advanced NLP would have been able to identify many of these cases.

It was important to specify for each item in the algorithm the range of dates to be included in the review, setting the time period (eg, days, weeks, months) before and after the date the item was identified in the EMR. Again using the example of RBC indices, our validation included evidence of prescription for several medications that might affect RBC indices, specifying a timing of 2 months before or after the traits were measured. However, we reviewed the entire record for presence of hereditary anemia, because this could be noted at any time in the EMR. Timing should also be considered for logic checks—for example, pregnancy is not expected in women over age 65 years, but was found in some records because of coding errors. When examining long periods, the algorithm must specify the minimum follow-up required and how to handle deaths and health plan disenrollment.

Repeated measures

Many phenotypes, for example, blood pressure and laboratory measures, are recorded repeatedly. This presents opportunities and challenges. The presence of multiple measures allows longitudinal studies such as progression of renal disease based on increase in serum creatinine over time. Repeated measures may also provide a more accurate representation of the trait than a single measure. Challenges to using these values include defining the relevant time period and specifying the most meaningful value to be studied (eg, overall median or mean, annual median or mean, age-adjusted median mean or median, change in value over time, highest value in each year).

Transience in the EMR

Individuals move in and out of medical systems or may be seen only for specialty care. Some institutions assign a lifetime identification number while others assign a new identification number with each episode of enrollment, making it difficult to link records across time, and resulting in incomplete or fragmented EMR data. Transience can have important repercussions for phenotype algorithms in the types of data elements used, the data sources interrogated, and the performance of the algorithm. Using hypothyroidism as an example, subjects were excluded based on a prior history of thyroidectomy as defined in the algorithm using diagnosis and procedure codes. However, validation at the Mayo Clinic identified several instances of thyroidectomy at another medical facility and thus not identified by the EMR algorithm. Algorithms may need to include specific considerations for enrollment as well as for patients who die during the study period. The Mayo Clinic site studied the role data fragmentation between medical centers played in identifying type 2 diabetics and found that using data from both medical centers improved both recall and precision.[7]

Review parameters

Scope of review

As EMR data accumulate, this issue is increasingly important. The scope of review can profoundly impact review time and thus project costs. Some factors (eg, evidence of cancer) may require review of the entire record, but for others (eg, chemotherapy receipt 1 year before or after a particular RBC value) a more reasonable and equally sound approach is to specify windows of interest. In the latter case, exclusions are applied at the sample level rather than the person level. Some variables will be absolute inclusions or exclusions regardless of when they occur, whereas others are applied to every repeated value (eg, blood pressure).

Combining research and EMR data

We usually chose to review only EMR data because these data are typically available when designing a study, and because the validation method was then consistent across eMERGE study sites. However, Group Health's participants were selected because they were enrolled in a longitudinal study of aging, and research data were available in addition to EMR data. Research data were far more detailed than EMR data because participants were seen in a research clinic every 2 years. We were thus able to use a research-quality dementia diagnosis to develop and validate the EMR-based phenotype algorithm. We found that ICD-9-CM code 331 had a positive predictive value of 79% when compared to a gold standard research-quality dementia diagnosis.[16]

Utility of pharmacy claims data

The Marshfield Clinic examined the relative contributions of insurance (claims) and EMR data for identifying the phenotype of resistant hypertension and controls without resistant hypertension. Subjects (n=3178) were selected from Marshfield Clinic's Personalized Medicine Research Project cohort who had

at least one primary care visit at Marshfield and had continuous insurance membership in Security Health Plan (a Marshfield Clinic owned HMO) from January 2005 through December 2009. Of the 3178 study subjects, 99.3% had at least one claim during the study period.

The resistant hypertension phenotype definition had two case groups[7]; only one could be evaluated using health plan data because blood pressure measurements were not available from the Security Health Plan data. The resistant hypertension definition that could be evaluated required the documented simultaneous use of four or more classes of blood pressure lowering medications on two separate occasions that were more than 1 month apart. Using both data sources, 32 subjects were identified, with 26 identified solely from the EMR, 5 identified solely from insurance claims, and 1 appearing in both data sources. Thus, using only EMR data would have reduced the number of identified cases by 15%, and using only insurance data would have reduced the case yield by 80%. However, since the insurance data source did not have blood pressure data, it could not be used for identifying cases using blood pressure measurements or for either of the two control definitions for resistant hypertension.

Validation steps

The value of iterative algorithm development

The development of phenotype pseudocode and phenotype validation worked best as iterative processes that involved informaticists, clinical content experts, epidemiologists, and geneticists. The value of validation went far beyond confirming that the phenotype was accurate. Information obtained at each step was used to fine-tune and improve the final phenotype algorithm and pseudocode (figure 1). The process had two phases. First, the primary site developed the pseudocode, which was reviewed by secondary sites, and then the process was tested at the primary site. In the second phase the phenotype was validated at secondary sites. Abstraction form development was also iterative, with one site drafting a form and all sites reviewing, giving input, pilot testing, and revising until the form was finalized. Making decisions about the validation process and conducting validation reviews is time-consuming. Pilot testing the algorithm or validation tool could require additional chart abstraction for each iteration (to avoid bias in final results). However, we found that the process was well worth the time and frequently identified unintended errors. Ultimately the time spent developing validation approaches contributed to more robust phenotype definitions.

Structured chart review versus physician review

Participating sites used one of two types of chart review. Some (Vanderbilt, Northwestern) used physicians to review charts for validation, with a written guide listing eligibility and exclusion criteria for cases and controls. Northwestern used two clinical researchers for chart review, with a physician reviewing results that differed between reviewers or from the outcome chosen by the pseudocode. Other sites (Group Health, Marshfield Clinic, and Mayo Clinic) developed chart abstraction forms based on the eligibility and exclusion criteria, and provided codebooks to define situations that might require interpretation. Marshfield also used clinical domain experts to assist with interpretation if the trained abstractors could not determine a specific status. Trained medical abstractors searched the clinical notes to record objective measurements and dates and interpreted the intent of the provider for the existence of a condition. Some sites reviewed the entire medical record (paper and EMR) while others reviewed only the electronic portion of the EMR (without paper records). Distinctions between materials reviewed depended in part on the amount of data available in the EMR at a particular institution.

Validation results

Once reviews were completed, we calculated the positive predictive value (PV+) for being a case (number of algorithm cases confirmed as true cases divided by the total number of algorithm cases), the PV+ for being a control (number of algorithm controls confirmed as true controls divided by the total number of algorithm controls), or the PV+ for meeting algorithm eligibility criteria (table 3). This approach was taken because we sought to identify for GWAS those persons who did and did not have the phenotype of interest or who met algorithm-derived

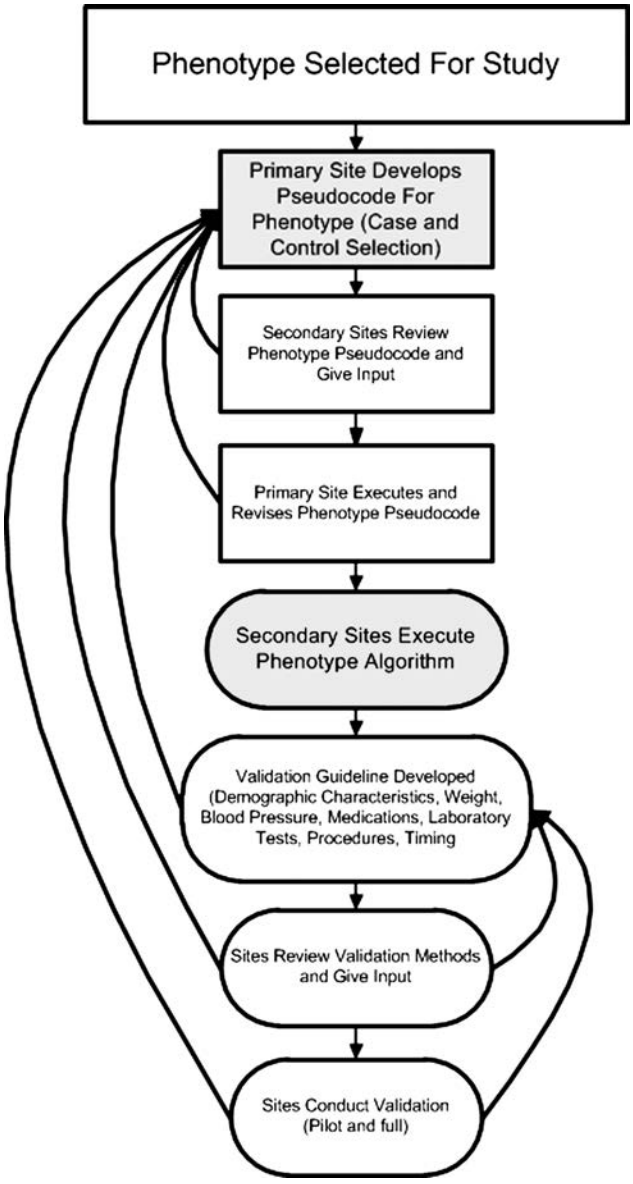

Figure 1 Phenotype development and validation. In this two-stage process a primary site first develops and executes the phenotype (boxes), and then secondary sites execute the phenotype (ovals). At each step feedback to primary and secondary sites may lead to revisions in the methods (arrows).

Table 3 Positive predictive value for phenotype case and control algorithms, and for phenotype eligibility algorithms across Electronic Medical Records and Genomics sites

Phenotype	Number validated	Positive predictive value				
		Group Health (%)	Marshfield Clinic (%)	Mayo Clinic (%)	Northwestern University (%)	Vanderbilt University (%)
Validated for case/control status and eligibility						
Cataract						
Case*	3234		97.7			96.0
Control*	3184		97.7			
Dementia						
Case*	3778	73.0	89.7			84.0
Control	505		96.7			
Type 2 diabetes						
Case	300		99.0		98.2	100
Control	143		98.0		100	100
Diabetic retinopathy						
Case	229		80.0			67.6
Control	80		98.0			100
Resistant hypertension						
Case	354	90.0	100		84.4	
Control	144	91.0	100		93.8	84.0
Peripheral arterial disease						
Case*	11504		87.5	90.7	95.0	
Control	100			100		
Chronic autoimmune hypothyroidism						
Case	389	92.0	91.3	82.0	98.1	98.0
Control	290	100	100	96.0	100	100
Validated for eligibility						
Low level of high density lipoprotein	440		81.6			
Lipids*	1054		78.8		92.3	
Red blood cell indices	391		96.4	98.0	98.0	
White blood cell indices	365		89.6		85.0	
QRS duration	245		100		96.9	97.0
Height	579		86.9		95.1	

Blank cells if did not participate in validation of that phenotype.
*Number large due to pre-existing study with validation.

eligibility criteria. Often, being a control was not simply the absence of being a case. For example, to be a control for primary hypothyroidism required the presence of a normal thyroid stimulating hormone test. Those without this test would meet neither the case nor the control criteria. For phenotypes such as height or laboratory values, which are easily obtained from the EMR, the critical point was the selection of subjects who met strict eligibility criteria. Our approach was thus unlike the validation of a simple test result where one is deemed positive or negative, and where sensitivity, specificity, predictive value positive and negative, and receiver operating curves are generated. Rather we validated separately our case and control definitions, and phenotype eligibility.

Most algorithms performed well, with PV+ values of 67.7–100%. Among 51 algorithm reviews across five sites, almost three quarters of the reviews yielded PV+ values of 90% or greater, and only three reviews yielded PV+ values less than 80%. For dementia, validation was poorer at Group Health because we compared research-quality, and research based ascertainment of dementia to the medical record. But even algorithms that performed less well at a particular site performed well overall.

DISCUSSION

Genetic research requires precise phenotype definitions, but EMR phenotype data is recorded inconsistently, in a variety of formats, and at times with biases. For example, blood pressure is recorded more frequently among people with hypertension, and people with chronic diseases have more frequent visits than those without them—both of which could lead to ascertainment bias. We observed great heterogeneity across the five EMRs of the eMERGE network sites, which included academic medical centers (Vanderbilt, Northwestern), health maintenance organizations (HMOs) (Group Health, Marshfield Clinic), and a large private health plan (Mayo Clinic). Network membership requires both EMR data and a large DNA biobank for genotyping. Phenotype algorithms across sites covered a variety of disease states and quantitative traits, and used billing codes, multi layer perceptron (MLP) structured diagnoses, medications, laboratory tests, and measures such as blood pressure, height, and weight. Despite different EMR infrastructures, we were able to develop and validate 13 diverse phenotypes, and algorithms typically performed well at each site tested. We have summarized our observations from this experience and offer specific recommendations for generation and validation of EMR phenotypes algorithms (table 4).

Table 4 Considerations and recommendations for phenotype validation

Issue	Considerations and recommendation
Validation approach	Have one site lead development of validation instructions and distribution to other sites for review before validation
Variables	
Selection of variables for validation	▶ Consider validating each variable in the phenotype definition or pseudocode: – Phenotype itself – Inclusion and exclusion criteria – Covariate information ▶ Consider variable consistency—validation is more important for subjective variables and variables that are particularly important to the analysis
Time frame for each variable	▶ Specify the time period of interest for each variable ▶ Be clear about intervals to be reviewed before and after the phenotype definition date
Repeated measures	▶ Define measures to be used ▶ Define measures consistently with phenotype definitions. For example, apply similar time restrictions
Transience in the EMR	When developing the validation instrument, consider that individuals move in and out of the system, or may enter only for specialty care. Some systems assign a new identification code with each enrollment, complicating longitudinal follow up. Individuals may have events (surgeries, diagnoses) before enrollment that are identifiable only through EMR review. Don't assume codes are sufficient
Review parameters	
Scope of review	▶ Remember that scope profoundly impacts the time to review a record ▶ Consider important time windows rather than reviewing the entire EMR
Combining research and EMR data	Be consistent: If a phenotype is selected from the EMR, use only the EMR for validation. Using additional sources such as more thorough research data can falsely raise the predictive power of the phenotype algorithm. Exceptions must be well justified
Review of claims versus medications dispensed	▶ Determine data provenance: Will medications be based on orders, fills or claims? ▶ In multisite studies all data may not be available at every site. Define acceptable gaps ▶ Determine how each source will be accessed for validation ▶ Decide if medications need to be validated
Validation steps	
Iteration	▶ Use an iterative process, prepared in advance (figure 1) ▶ Refine phenotype definitions and improve through validation ▶ Choose more informative multisite validation over single-site validation
Type of validation (content expert vs structured chart review)	▶ Either can be used, but strive for consistency in the two approaches when both are used ▶ Make guidelines for content expert reviews as specific as possible ▶ Simultaneously develop content expert and structured review guidelines

EMR, electronic medical record.

A major effort of eMERGE has been generating and validating electronic phenotype algorithms. Table 3 demonstrates that the majority of these algorithms ported well to diverse sites. These data confirm what has been shown in prior eMERGE and Pharmacogenomics Research Network studies regarding algorithm transportability in primary hypothyroidism,[12] type 2 diabetes,[18] cataracts,[19] and rheumatoid arthritis.[20] It is important to note that these evaluations cross different EMR implementations, different NLP systems, and different fundamental types of algorithms, from deterministic to logistic regression. eMERGE algorithms are posted on PheKB.org, which hosts the original versions and implementation data for completed algorithms. As other sites deploy and evaluate algorithms, other users can post this data as well.

While an EMR increases efficiency, using it correctly requires effort. One approach to streamline phenotype definitions might be structures that facilitate creation of algorithms using standard terminologies. By representing covariates and algorithm components with standard terminologies, developers increase the ease with which algorithms can be compared and, potentially, reused. This is especially true for the outputs and covariates resulting from them. One such tool is eleMAP, which was developed by eMERGE investigators. This free, online tool allows researchers to harmonize their local phenotype data dictionaries to existing metadata and terminology standards such as the caDSR (Cancer Data Standards Registry and Repository), NCIT (NCI Thesaurus), and SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms). eleMAP can be used to search and browse metadata related to different studies, create new studies (and the related metadata), and export metadata in Microsoft Excel format.[21]

Proposed data infrastructures, such as PhenX,[22] offer catalogs of standard measures to be used in GWAS. The adoption of such standard measures might, in principle, have made the eMERGE studies and phenotype validation easier. However, the success of common data infrastructures will require communication between parties using EMRs for different purposes. Research needs are not the highest priority for those developing and using EMRs, and standards such as PhenX are often proposed with researchers rather than clinicians in mind. The evolving standards for interoperability that are part of the Health Information Technology for Economic and Clinical Health (HITECH) Act may improve EMR data representation and quality.

Another approach that can streamline phenotype definition is demonstrated by the HMO Research Network's (HMORN) Virtual Data Warehouse (VDW).[23] The HMORN, of which Group Health and Marshfield Clinic are members, is an organization of 19 HMO-based research programs whose mission is to use their collective capabilities to integrate research and practice to improve health and healthcare. The VDW consists of parallel databases set up identically at each HMORN site that can be merged across sites. The databases were constructed by extracting data directly from the local electronic data systems

and reconfiguring them to use standard variable names and coded values. A project analyst writes a program based on the VDW data dictionary that is sent to participating sites to be run locally, and output files are transferred securely. We have found that even with this resource, validating and confirming VDW components that are subject to practice variation (eg, ICD coding choices, prescriptions) is critical.

Efforts to use EMR data for research depend on the ability of heathcare institutions to establish and maintain an effective EMR.[2] This requires a team with expertise in technology, clinical, process redesign, management, and informatics. Resources for standardizing collections of clinical data are available from the federal government. For example, the Surgeon General's Family Health History Initiative 2011[24] provides a simple, web-based tool for patients to enter family history in a standardized format, for inclusion in an EMR. The Office of the National Coordinator for Health Information Technology provides information and assistance on meaningful EMR use, including coding standards for key data elements.

EMRs cannot capture all nuances of patient–provider interactions, but they are extremely useful resources for well designed, informative clinical studies. Accurate EMR capture of diagnosis, laboratory, and medication data, supplemented with text-mining tools and NLP, can provide excellent phenotype data for genomic studies, including GWAS. However, even with advances and new approaches, the heterogeneity in EMRs means that phenotype validation will remain an important aspect of their use.

Our approach had limitations. Ideally, validation methods would be cross-validated using external reviewers to ensure consistent phenotype assessment. This cross-validation was beyond our resources, and could create challenges with local IRBs and system access restrictions. However, using standard validation forms and processes may serve as a surrogate to reduce variation. A second limitation was the use of both expert physician reviewers with written guidelines and medical chart abstractors with a structured abstraction form. Our belief is that such a combination of review techniques may actually be a strength; reviewing physicians may note factors that exclude a person as a case not envisioned in a chart abstraction, and formal chart abstraction may identify logical inconsistencies with a more meticulous review. Though we did not formally evaluate whether these two approaches gave different validation answers, no sizeable differences in validation outcomes were seen across sites.

CONCLUSIONS

Despite the diverse structure of the five EMRs of the eMERGE sites, we developed, validated, and successfully deployed 13 electronic phenotype algorithms. Validation is a worthwhile process that not only measures phenotype performance but also strengthens phenotype algorithms and enhances their inter-institutional sharing.

Contributors KMN took the lead to draft the manuscript and takes responsibility for its contents. All authors provided input to the study design at their respective sites and the overall network objectives, and all authors read and approved the final manuscript.

Funding The eMERGE Network was initiated and funded by NHGRI, with additional funding from the National Institute of General Medical Science 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 Genetic Medicine.

Competing interests None.

Ethics approval All active sites approved the study.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Office of the National Coordinator for Health Information Technology. Electronic Health Records and Meaningful Use. 2011; http://healthit.hhs.gov/portal/server.pt?open=512&objID=1325&parentname=CommunityPage&parentid=1&mode=2. (accessed 31 May 2011).
2 Walker JM. Electronic medical records and health care transformation. *Health Aff (Millwood).* 2005;24:1118–20.
3 Walker J, Pan E, Johnston D, Adler-Milstein J, *et al*. The value of health care information exchange and interoperability. *Health Affairs*, no. (2005): doi:10.1377/hlthaff.w5.10 (accessed 18 May 2013).
4 Edwards BJ, Haynes C, Levenstien MA, *et al*. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genetics* 2005;6:18 doi:10.1186/1471-2156-6-18 (accessed 18 May 2013).
5 Rice JP, Saccone NL, Rasmussen E. Definition of the phenotype. *Adv Genet* 2001;42:69–76.
6 Burton PR, Hansell AL, Fortier I, *et al*. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2009;38:263–73.
7 National Human Genetics Research Institute. The eMERGE Network Electronic Medical Records and Genomics. 2011; http://www.gwas.net (accessed 31 May 2011).
8 Kullo I, Ding K, Jouni H, *et al*. A genome-wide association study of red blood cell traits using the electronic medical record. *PLoS ONE* 2010;5(9):e13011. doi:10.1371/journal.pone.0013011 (accessed 18 Mar 2013).
9 Crosslin DR, McDavid A, Weston N, *et al*. Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. *Hum Genet* 2012;131:639–52.
10 Denny JC, Ritchie MD, Crawford DC, *et al*. Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. *Circulation* 2010;122:2016–21.
11 Kullo IJ, Ding K, Shameer K, *et al*. Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. *Am J Hum Genet* 2011;89:131–8.
12 Denny JC, Crawford DC, Ritchie MD, *et al*. 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* 2011;89:529–42.
13 Wojczynski MK, Tiwari HK. Definition of phenotype. *Adv Genet* 2008;60:75–105.
14 Kullo IJ, Fan J, Pathak J, *et al*. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J Am Med Inform Assoc* 2010;17:568–74.
15 McCarty CA, Chisholm RL, Chute CG, *et al*. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011;4:13. doi:10.1186/1755-8794-4-13 (accessed 18 Mar 2013).
16 Kho AN, Pacheco JA, Peissig PL, *et al*. Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. *Sci Transl Med* 2011;3(79):79re1. doi:10.1126/scitranslmed.3001807 (accessed 18 Mar 2013).
17 Conway M, Berg RL, Carrell D, *et al*. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. *AMIA Annu Symp Proc* 2011;2011:274–83.
18 Kho AN, Hayes MG, Rasmussen-Torvik L, *et al*. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2012;19:212–18.
19 Peissig PL, Rasmussen LV, Berg RL, *et al*. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *J Am Med Inform Assoc* 2012;19:225–34.
20 Carroll RJ, Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012;19(e1):e162–9. Epub 2012 Feb 28 (accessed 18 Mar 2013).
21 Pathak J, Wang J, Kashyap S, *et al*. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J Am Med Inform Assoc* 2011;18:376–86.
22 Hamilton CM, Strader LC, Pratt JG, *et al*. The PhenX Toolkit: get the most from your measures. *Am J Epidemiol* 2011;174:253–60.
23 Hornbrook MC, Hart G, Ellis JL, *et al*. Building a virtual cancer research organization. *J Natl Cancer Inst Monogr.* 2005:12–25.
24 US Department of Health and Human Services. Surgeon General's Family Health Initiative. 2011; http://www.hhs.gov/familyhistory/ (accessed 31 May 2011).