Electronic Medical Records as a Tool in Clinical Pharmacology: Opportunities and Challenges

DM Roden[1,2], **H Xu**[3], **JC Denny**[1,3], and **RA Wilke**[1]

[1]Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA

[2]Department of Pharmacology, Vanderbilt University School of Medicine, Nashville, Tennessee, USA

[3]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA

Abstract

The development and increasing sophistication of electronic medical record (EMR) systems hold the promise of not only improving patient care but also providing unprecedented opportunities for discovery in the fields of basic, translational, and implementation sciences. Clinical pharmacology research in the EMR environment has only recently started to become a reality, with EMRs becoming increasingly populated, methods to mine drug response and other phenotypes becoming more sophisticated, and links being established with DNA repositories.

EMR CAPABILITIES

The 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act is providing substantial economic incentives for health practitioners to adopt EMR systems, by offering incentive payments for "meaningful use" of EMRs and (scheduled to begin in 2015) decreasing Medicare payments to practitioners who do not adopt such standards.[1] As these capabilities become more widely embedded across the health-care system, there is an increasing opportunity for generating very large sets by developing collaborations among health care systems with EMRs, e.g., the HMO Research Network, the Agency for Healthcare Research and Quality–funded Scalable Partnering Network (SPAN), and the electronic Medical Records and Genomic (eMERGE) Network of the National Human Genome Research Institute.[2–4]

EMRs can include a range of capabilities that in turn affect how useful they can be in medical practice. At the simplest level, an EMR is a longitudinal record of the health status of an individual patient, leading to increased efficiencies through simple changes in practice, such as more rapid and simultaneous access to medical information and reducing costs associated with paper storage and retrieval. EMR systems receive patient data directly from users and also from diverse subsystems (such as pharmacies and radiology records) through common protocol languages (typically Health Language 7) and serve as an information repository for more advanced patient-specific guidance through computerized provider order entry and electronic prescribing systems as practitioners execute orders for individual patients. Such capabilities can provide reminders of the required preventive medicine alerts,[5]

Correspondence: DM Roden (dan.roden@vanderbilt.edu).

CONFLICT OF INTEREST
The authors declared no conflict of interest.

potential drug–drug interactions,[6] and guidance for situations in which laboratory testing can be avoided.[7] EMRs can provide tools to aggregate outcomes for many patients and thereby provide practices with metrics such as drug utilization, adherence to guidelines, and real-time surveillance of at-risk populations.[8] An extension of this concept is to aggregate information across EMR systems; for example, the implementation of a Regional Health Information Exchange across emergency rooms in Memphis, Tennessee, decreased utilization of computed tomography scans and laboratory tests and resulted in fewer hospital admissions.[9] Ultimately, EMR systems can become central elements in large-scale public health surveillance. For example, a retrospective examination of myocardial infarction after the introduction of rofecoxib and celecoxib revealed an ~30% increase in myocardial infarction rates that returned to baseline after the removal of rofecoxib from the market in 2004.[10] The idea that automated surveillance systems can provide an early warning of such events is one rationale for the Sentinel initiative of the US Food and Drug Administration.[11]

Notably, EMRs can be used for discoveries of this type, in parallel with other sources of such information: tracking Google searches allowed investigators to detect the onset of the 2008 influenza epidemic with a one-day time lag, far shorter than could be achieved with conventional surveillance.[12] Another database that has been analyzed in parallel with EMR data is the FDA's Adverse Event Reporting System. Tatonetti and colleagues[13] developed methods to identify a profile (often involving drug exposures or diagnoses) of patients with abnormalities of glucose homeostasis and deployed these in the Adverse Event Reporting System. They used their definition to screen for pairs of drugs that individually had no effect on glucose homeostasis but in combination exerted a strong effect. This approach implicated the combination of pravastatin and paroxetine in promoting hyperglycemia, and the same signal was then sought in three EMR systems. Neither drug by itself affected blood glucose levels. However, in 104 patients with diabetes and 135 without diabetes who had received the combination, blood glucose concentration increased 19 mg/dl overall, and in those with diabetes the change was even larger, 48 mg/dl. Further study indicated that this is not a general effect of combined therapy with selective serotonin reuptake inhibitors and statins.

The idea of coupling DNA repositories to EMR systems[14,15] raises the possibility, now validated across multiple studies as described below, that the EMR can be used in genomic and pharmacogenomic research, e.g., to replicate or discover genotype–phenotype associations. One common model is to seek out and prospectively obtain consent from volunteers for such projects.[14] An advantage of such repositories is the ability to recontact patients (e.g., to obtain further phenotypic information). An alternative approach is to extract DNA from discarded tissue and couple information on genomic variation to de-identified EMR information.[15] Natural-language text de-identification systems[16] are used to suppress personal health information while retaining the richness of clinical content necessary to determine whether an individual meets the required criteria. The advantage of this approach, which can be coupled to an "opt-out" capability,[15,17] is that large cohorts can be accrued relatively quickly and inexpensively. A disadvantage is that it does not allow recontacting patients, so that whatever analysis is proposed must rely exclusively on information contained in the EMR.

EMR-BASED DISCOVERY IN CLINICAL PHARMACOLOGY

Approach to EMR-based phenotyping

A common experimental design starts with construction of a cohort of cases and controls for a specific phenotype (e.g., a disease or exposure status to a medication) from EMR data. A manual chart review—the traditional approach for phenotyping— has been successfully used in an EMR environment, but it is cumbersome and time-consuming and generally cannot generate very large cohorts.[18] More recent efforts have been focused on developing

electronic algorithms for determining specific phenotypes from EMRs.[19,20] The criterion that we and others have adopted[2] is that algorithms should have very high positive predictive values (PPVs, generally >90–95%) to be able to identify cohorts for case–control studies. The general approach has been to develop an algorithm and deploy it in an EMR system until a set of cases (generally 50–100) is identified for manual review. This manual review then determines the PPV and deficiencies in the electronic phenotyping algorithm. The algorithm is then refined and the process iterated until the threshold PPV is attained. Especially with respect to common diseases, a similar process of algorithm development with validation and iterative refinement is undertaken to select control cohorts.

In general, the use of simple coded data such as International Classification of Disease (ICD-9) codes alone generates PPVs that do not approach the >95% PPV criterion.[20] Accordingly, phenotype algorithms often consider both structured (e.g., ICD-9 codes or lab results) and narrative (e.g., various types of clinical notes, text messages between patients and care providers) data in EMRs.[19] Natural-language processing technologies that can extract structured information (e.g., smoker: yes or no) from unstructured narrative clinical text have been used in such algorithms.[21] The use of natural-language processing is especially important to detect diseases and events occurring at facilities outside the recording center and entered as part of the patient's past medical history and for discovery of rare events that may not be represented in typical coding systems. As described below, well-defined electronic phenotyping algorithms have been successfully developed and employed in EMR-based studies of variability in disease presentations and genetic associations.[19,22,23]

One appealing feature of EMRs is that their inclusion of longitudinal data should enable studies of variability in phenotype such as those associated with disease complications, the tempo of disease progression, and response to drug exposures. Accurate identification of drug exposures from longitudinal observational EMR data is a key example of the challenges in implementing this idea. Although various types of structured medication data (e.g., physician order entries and drug administration records) may be available in inpatient settings, much of the information related to outpatient drug exposures resides in the narrative text of clinic notes or interoffice communications (e.g., phone call records with patients) and may therefore not be readily available for data analysis. Drug exposures change frequently over time because of intolerance, insurance status, patient/provider preferences, and the desire to reach clinical targets. Compliance is another important issue in drug-related studies. The manual method, in which a domain expert reviews all the different sources to compile a longitudinal history of drug uses for identifying drug exposure, is cumbersome and costly and often results in study cohorts with limited sample size.[18] As with disease definitions, advanced informatics methods that combine multiple sources of data have been developed.[24] These methods have successfully exploited the longitudinal nature of the EMR to identify drug–response phenotypes such as those associated with cardiovascular events during clopidogrel therapy after coronary stenting.[25] More sophisticated drug–response phenotype definitions, such as international normalized ratio prolongation during warfarin therapy[26] and the development of dose–response curves by analysis of low-density lipoprotein responses to multiple statin doses,[27] have been successfully developed. Through the rigorous characterization of phenotype, misclassification bias is minimized, thereby yielding greater insights into the true genetic architecture underlying treatment response.

A key lesson in these efforts is that high-quality phenotypes are best generated by collaboration among investigators with expertise in multiple disciplines, including biomedical informatics, epidemiology, and clinical pharmacology. It is especially important to engage practitioners in the specific disease domains being studied: these individuals have specific expertise on how diseases are represented in areas such as coding, laboratory values,

and free text and how diseases with similar presentations (e.g., rheumatoid arthritis vs. systemic lupus erythematosus) are distinguished in the EMR environment. Indeed, one advantage of the inclusion of longitudinal data in EMR-based phenotyping is the ability to distinguish among related diagnoses such as these on the basis of data from the multiple visits represented in an EMR. Preliminary data indicate that phenotype definitions developed in one EMR can be successfully deployed in others.[28,29]

Gene–disease associations

The genome-wide association study (GWAS) paradigm has been successfully applied to hundreds of human traits, and common single-nucleotide polymorphisms (SNPs) associated with these phenotypes have been identified and often replicated in large community-based or research-purposed cohorts. One of the first studies[19] to demonstrate that these associations can be replicated in an EMR environment analyzed 21 SNPs that had previously been implicated as common variants predisposing to atrial fibrillation, Crohn's disease, multiple sclerosis, rheumatoid arthritis, or type 2 diabetes. For each phenotype, natural-language processing techniques and billing code queries were applied to a set of 9,483 samples to identify cases ($n = 70$–698) and controls ($n = 808$–3,818). Each of the 21 tests of association yielded point estimates in the same direction as reported in the literature, and a majority of genotype–phenotype associations with reported odds ratios >1.25 were replicated at the $P < 0.05$ level.

Nodes in the eMERGE network have developed several algorithms to identify common phenotypes and have deployed these across several EMR systems[29] to replicate known GWAS-derived associations and identify new ones.[28,30,31] One report highlighted the advantage of performing studies in EMRs by repurposing GWAS-level genotyping data previously obtained for other phenotypes so as to identify a potential association between SNPs near *FOXE1*, a transcription factor previously associated with thyroid cancer and primary hypothyroidism.[28]

Application to pharmacogenomics

It is only in the past several years that EMRs have been coupled to DNA repositories and methods developed to identify cases and controls for disease association studies. The use of these resources for pharmacogenomics has lagged further behind, in part because the phenotyping for drug response is, as discussed above, necessarily more complex, with the numbers of cases being smaller. Nevertheless, studies have been conducted to examine drug–response phenotypes that can be readily extracted from EMRs with respect to many commonly used drugs.

HMG-CoA reductase inhibitors (statins) are the most commonly prescribed class of drugs in developed countries, and much of the observed variability in the low-density lipoprotein cholesterol–lowering efficacy of statins can be attributed to underlying genetic factors.[32] Yet GWAS published by our group and others have resolved very few loci associated with statin-induced change in low-density lipoprotein cholesterol at a level reaching genome-wide significance.[33] This is in part because treatment trials often contain data for only a single dose. By applying natural-language processing algorithms to data contained within EMRs, mathematical parameters representing potency (ED_{50}) and efficacy (E_{max}) can be derived from full dose–response curves, thereby enabling the world's growing biobanks to overcome the initial obstacles faced during early statin GWAS.

Estrogen-receptor SNPs have been implicated as risk factors for venous thromboembolic disease during tamoxifen treatment in an EMR environment.[34] The association between *CYP2C19*2* and cardiovascular risk in patients receiving clopidogrel after coronary stenting

was replicated in an EMR environment, and an association with a variant in *ABCB1*, identified in some but not all previous studies, was also observed.[25] Similarly, associations between variants in *CYP2C9, VKORC1*, and *CYP4F2* and steady-state warfarin dose requirement in Caucasians were replicated in the EMRs, and new associations were identified with different *CYP2C9* variants as well as with variants in *CALU* in African Americans.[35] The clopidogrel and warfarin experiments leveraged significant human review to refine extraction of dose and assess compliance; the warfarin association was later replicated using only informatics techniques, yielding nearly the same results.[26] EMRs coupled to DNA repositories may also be a resource for studying the genomics of rare drug responses, although this vision has not yet come to fruition.[36]

Phenome scanning

The inclusion of many and diverse diagnoses per record in EMR systems that also include extensive genomic data opens up the possibility of phenome scanning.[37,38] The concept can be thought of as the inverse of the GWAS paradigm; a phenome- wide association study (PheWAS) examines the relationship between a single genetic variant and a large range of phenotypes,[38] currently represented as a series of ICD-9 codes. Preliminary studies demonstrated that the PheWAS method could replicate known phenotype–genotype associations.[38] The eMERGE hypothyroidism GWAS examined 1,317 cases and 5,053 controls; a parallel PheWAS in 13,617 patients readily replicated the association with *FOXE1* SNPs (with a *P* value $<10^{-13}$) and identified additional associations with thyroiditis, nodular and multinodular goiters, and thyrotoxicosis, but not with Graves' disease.[28]

SUMMARY: CHALLENGES AND OPPORTUNITIES

This description should make clear that the potential is now being realized for EMRs to act as a resource for discovery of new drug actions and of genomic influences on disease phenotypes and drug responses. The need to develop large data sets to study variability in human traits, including drug responses, can be met by single EMRs—in particular, by coupling EMRs to networks. Phenotyping presents a challenge; although the EMR is, by its nature, a "messy" (noisy) data set, methods to identify valid cases and controls are being refined. In addition, these studies can point to new types of information that might be desirable to include in the EMR. Examples include structured family histories and improved methods to document drug exposures and to couple pharmacy records to EMRs. Charts may be fragmented, and records of care delivered may be distributed across many EMR systems; interestingly, the Veterans Affairs EMR is an example of an integrated system that can deliver information at many nodes nationally. Although studies in the EMR environment raise privacy concerns, the EMR also provides a platform for examining the extent to which individual identities can be protected.[39]

A widely held vision is that information on genomic variation will be used to inform decision making for prevention, prognosis, and treatment. The discipline of pharmacogenomics is identifying an increasing number of variants associated with drug responses; however, the very success of these efforts represents a barrier to implementation because no human can be expected to keep track of this increasing data set and its implications for drug prescribing. The capability of advanced EMR systems to archive large amounts of individual data and deliver advice to providers at the point of care seems to offer an obvious solution to this problem. Implementing this concept presents many challenges; these are further outlined in our recent description of the Vanderbilt PREDICT project, a program that deposits genomic variant data into EMRs on a preemptive basis and delivers point-of-care electronic decision support if a drug with known pharmacogenomic variability is prescribed.[40]

Acknowledgments

References

1. Blumenthal D. Wiring the health system—origins and provisions of a new federal program. N Engl J Med. 2011; 365:2323–2329. [PubMed: 22168647]

2. McCarty CA, 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. [PubMed: 21269473]

3. Toh S, Platt R, Steiner JF, Brown JS. Comparative-effectiveness research in distributed health data networks. Clin Pharmacol Ther. 2011; 90:883–887. [PubMed: 22030567]

4. Platt R, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. Pharmacoepidemiol Drug Saf. 2001; 10:373–377. [PubMed: 11802579]

5. Dexter PR, Perkins SM, Maharry KS, Jones K, McDonald CJ. Inpatient computer-based standing orders vs physician reminders to increase influenza and pneumococcal vaccination rates: a randomized trial. JAMA. 2004; 292:2366–2371. [PubMed: 15547164]

6. Ko Y, et al. Practitioners' views on computerized drug–drug interaction alerts in the VA system. J Am Med Inform Assoc. 2007; 14:56–64. [PubMed: 17068346]

7. Neilson EG, et al. The impact of peer management on test-ordering behavior. Ann Intern Med. 2004; 141:196–204. [PubMed: 15289216]

8. Waitman LR, et al. Adopting real-time surveillance dashboards as a component of an enterprisewide medication safety strategy. Jt Comm J Qual Patient Saf. 2011; 37:326–332. [PubMed: 21819031]

9. Frisse ME, et al. The financial impact of health information exchange on emergency department care. J Am Med Inform Assoc. 2012; 19:328e333. [PubMed: 22058169]

10. Brownstein JS, Sordo M, Kohane IS, Mandl KD. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. PLoS ONE. 2007; 2:e840. [PubMed: 17786211]

11. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System—a national resource for evidence development. N Engl J Med. 2011; 364:498–499. [PubMed: 21226658]

12. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457:1012–1014. [PubMed: 19020500]

13. Tatonetti NP, et al. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. Clin Pharmacol Ther. 2011; 90:133–142. [PubMed: 21613990]

14. McCarty CA, Nair A, Austin DM, Giampietro PF. Informed consent and subject motivation to participate in a large, population-based genomics study: the Marshfield Clinic Personalized Medicine Research Project. Community Genet. 2007; 10:2–9. [PubMed: 17167244]

15. Roden DM, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008; 84:362–369. [PubMed: 18500243]

16. Aberdeen J, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform. 2010; 79:849–859. [PubMed: 20951082]

17. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. Principles of human subjects protections applied in an opt-out, de-identified biobank. Clin Transl Sci. 2010; 3:42–48. [PubMed: 20443953]

18. Wilke RA, Berg RL, Vidaillet HJ, Caldwell MD, Burmester JK, Hillman MA. Impact of age, CYP2C9 genotype and concomitant medication on the rate of rise for prothrombin time during the first 30 days of warfarin therapy. Clin Med Res. 2005; 3:207–213. [PubMed: 16303885]

19. Ritchie MD, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet. 2010; 86:560–572. [PubMed: 20362271]

20. Wilke RA, et al. Use of an electronic medical record for the identification of research subjects with diabetes mellitus. Clin Med Res. 2007; 5:1–7. [PubMed: 17456828]

21. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128–144. [PubMed: 18660887]

22. Kurreeman F, et al. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. Am J Hum Genet. 2011; 88:57–69. [PubMed: 21211616]

23. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011; 12:417–428. [PubMed: 21587298]

24. Liu M, et al. Modeling drug exposure data in electronic medical records: an application to warfarin. AMIA Annu Symp Proc. 2011; 2011:815–823. [PubMed: 22195139]

25. Delaney JT, et al. Predicting clopidogrel response using DNA samples linked to an electronic health record. Clin Pharmacol Ther. 2012; 91:257–263. [PubMed: 22190063]

26. Xu H, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. J Am Med Inform Assoc. 2011; 18:387–391. [PubMed: 21672908]

27. Peissig P, et al. Construction of atorvastatin dose-response relationships using data from a large population-based DNA biobank. Basic Clin Pharmacol Toxicol. 2007; 100:286–288. [PubMed: 17371534]

28. Denny JC, 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; 89:529–542. [PubMed: 21981779]

29. Kho AN, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med. 2011; 3:79re1.

30. Crosslin DR, et al. Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. Hum Genet. 2012; 131:639–652. [PubMed: 22037903]

31. Kullo IJ, et al. Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. Am J Hum Genet. 2011; 89:131–138. [PubMed: 21700265]

32. Mangravite LM, et al. Combined influence of LDLR and HMGCR sequence variation on lipid-lowering response to simvastatin. Arterioscler Thromb Vasc Biol. 2010; 30:1485–1492. [PubMed: 20413733]

33. Thompson JF, et al. Comprehensive whole-genome and candidate gene analysis for response to statin therapy in the Treating to New Targets (TNT) cohort. Circ Cardiovasc Genet. 2009; 2:173–181. [PubMed: 20031582]

34. Onitilo AA, et al. Estrogen receptor genotype is associated with risk of venous thromboembolism during tamoxifen therapy. Breast Cancer Res Treat. 2009; 115:643–650. [PubMed: 19082882]

35. Ramirez AH, et al. Predicting warfarin dosage in European-American and African-American subjects using DNA samples linked to an electronic health record. Pharmagenomics. 2012; 13:407–418.

36. Wilke RA, et al. Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. Nat Rev Drug Discov. 2007; 6:904–916. [PubMed: 17971785]

37. Ghebranious N, McCarty CA, Wilke RA. Clinical phenome scanning. Pers Med. 2007; 4:175–182.

38. Denny JC, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010; 26:1205–1210. [PubMed: 20335276]

39. Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. Proc Natl Acad Sci USA. 2010; 107:7898–7903. [PubMed: 20385806]

40. Pulley JM, et al. Operational implementation of prospective genotyping for personalized medicine: the design of the Vanderbilt PREDICT project. Clin Pharmacol Ther. (in press).