The Emerging Role of Electronic Medical Records in Pharmacogenomics

RA Wilke[1], H Xu[2], JC Denny[2], DM Roden[1], RM Krauss[3], CA McCarty[4], RL Davis[5], T Skaar[6], J Lamba[7] and G Savova[8,9,10]

Health-care information technology and genotyping technology are both advancing rapidly, creating new opportunities for medical and scientific discovery. The convergence of these two technologies is now facilitating genetic association studies of unprecedented size within the context of routine clinical care. As a result, the medical community will soon be presented with a number of novel opportunities to bring functional genomics to the bedside in the area of pharmacotherapy. By linking biological material to comprehensive medical records, large multi-institutional biobanks are now poised to advance the field of pharmacogenomics through three distinct mechanisms: (i) retrospective assessment of previously known findings in a clinical practice-based setting, (ii) discovery of new associations in huge observational cohorts, and (iii) prospective application in a setting capable of providing real-time decision support. This review explores each of these translational mechanisms within a historical framework.

Although the field of pharmacogenomics has the potential to transform the clinical practice of medicine, gene-based drug prescribing currently occurs rather infrequently. The Clinical Pharmacogenomics Implementation Consortium has been organized to move research findings into routine practice (http://www.PharmGKB.org). Personalized gene-based health-care delivery may therefore soon shift the allocation of medical resources away from reactive treatment of disease toward a more proactive approach based upon interindividual variations.[1,2] It is likely that electronic medical records (EMRs) will play a pivotal role in this transformation.

Before 2008, only ~10% of all US physicians were using a basic electronic health-care record.[3] Although the proportion of US hospitals implementing a basic EMR system was similarly low (7.6%) before 2008,[4] the application of EMRs has subsequently been expanding rapidly in both the inpatient and outpatient settings.[5] Nearly half of all large multispecialty group practices now utilize a comprehensive electronic record (http://www.computerworld.com).

Through the Patient Protection and Affordable Care Act of 2010, federal legislators have set an aggressive timeline to encourage the widespread implementation of EMRs.[5] Providers who roll out EMRs by 2015 will receive additional incentives for the care of Medicare patients, but only if the application is determined to fulfill prespecified "meaningful use" criteria. The meaningful use rules for 2011 and 2012 were released by the US Department of Health and Human Services on 13 July 2010 (http://www.ama-assn.org). A two-track approach has been established, containing criteria that are both mandatory (e.g., maintenance of active medication lists) and optional (e.g., decision-support software capable of flagging drug–drug interactions).

To date, EMR deployment not only has improved patient care, but has also established large, practice-based longitudinal data sets ideal for the conduct of observational research.[6,7] These data sets are rich in clinical information and available in both structured and unstructured formats. Structured data include diagnoses, clinical laboratory results, diagnostic imaging results, procedures ordered and performed, medications ordered and dispensed, and physician order entry. Structured medication data, in particular, have been used for a large variety of pharmacoepidemiology, pharmacoeconomic, and service-related health-care investigations.[8]

[1]Department of Medicine, Division of Clinical Pharmacology, Vanderbilt University Medical Center, Nashville, Tennessee, USA; [2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA; [3]Department of Atherosclerosis Research, Center for Nutrition and Metabolism, Children's Hospital Oakland Research Institute, Oakland, California, USA; [4]Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, USA; [5]Center for Health Research Southeast, Kaiser Permanente, Atlanta, Georgia, USA; [6]Department of Medicine, Division of Clinical Pharmacology, Indiana University School of Medicine, Indianapolis, Indiana, USA; [7]Department of Experimental and Clinical Pharmacology, University of Minnesota, Minneapolis, Minnesota, USA; [8]Department of Biomedical Informatics, Mayo Clinic, Rochester, Minnesota, USA; [9]Children's Hospital Informatics Program, Children's Hospital Boston, Boston, Massachusetts, USA; [10]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. Correspondence: RA Wilke (russell.a.wilke@vanderbilt.edu)

Received 26 June 2010; accepted 30 September 2010; advance online publication 19 January 2011. doi:10.1038/clpt.2010.260

When unstructured medication data are embedded in free-text clinical notes, natural language processing (NLP) algorithms have proven useful in the accurate reconstruction of comprehensive drug-exposure histories.[9] NLP-based phenotyping approaches have been used successfully to identify genetic determinants of drug outcome in the context of toxicity and efficacy.[10]

As a community, our ability to characterize the genetic architecture underlying treatment outcomes also continues to improve because of advances in genotyping technology. Increases in throughput and decreases in cost are allowing investigators to move from searches for candidate genes (pharmacogenetics) to genome-wide SNP scanning (pharmacogenomics) in cohorts of increasing size.[11] It is likely that entire genomic sequences will soon be linked to individual EMRs.[1,2] Because the convergence of these two rapidly expanding technologies (i.e., biomedical informatics and high-throughput genotyping) represents an unprecedented opportunity to bring functional genomics to the bedside, many large medical centers are constructing DNA biobanks in the context of routine clinical practice.[12–14] These biobanks offer the advantages of scale, cost efficiency, and extremely dense longitudinal health-care data.

Many models have emerged for the construction of biobanks, including models based on recruitment and enrollment of subjects from a specific geographic region or practice community. Other approaches have employed novel informatics strategies for total de-identification of EMRs. The latter approach optimizes security while allowing the de-identified samples to be linked to archived biological material in a cost-effective manner (as depicted in Figure 1). Early results indicate that biobanks constructed from completely de-identified EMRs are robust in their ability to replicate genetic associations previously identified in disease-specific cohorts.[15]

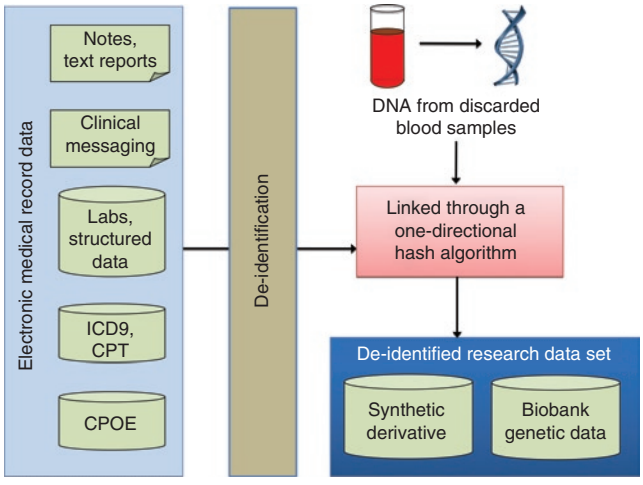

Figure 1 One approach to the construction of a biobank for pharmacogenomic research. Electronic medical records (EMRs) typically contain a combination of unstructured text reports and structured data. Structured data include most laboratory values, vital signs, and information such as computerized provider order entry (CPOE) records. In addition, administrative billing codes (ICD9, CPT) form valuable components for electronic phenotyping. These data can then be de-identified using algorithms to remove personal health identifiers from text through a combination of statistical and pattern-matching techniques.[13] Finally, de-identified medical records are linked to DNA samples, using research-unique identifiers that can be generated using a one-way hash algorithm that prevents discovery of the input number (e.g., a medical record number). CPT, current procedural terminology; ICD, International Classification of Disease.

This review explores potential methods through which existing EMRs may help facilitate the translation of pharmacogenomics into widespread clinical practice. Three distinct mechanisms are discussed: retrospective assessment of known findings in a clinical practice-based setting, discovery of novel associations in the context of gene–environment interaction, and prospective application in a setting capable of providing real-time decision support through bioinformatics and pharmacovigilance. Each is presented in a historical context with an emphasis on future outlook.

HISTORICAL OVERVIEW

The clinical practice community began moving toward implementation of EMRs nearly half a century ago.[16] Although most early EMRs consisted primarily of billing codes, some contained diagnostic codes in site-specific lexicons. The majority of the early lexicons were rather cursory. Shortly thereafter, procedural codes and clinical laboratory data began to be archived in coded and easily extractable formats.[17] In most systems of care, clinical notes were entered as free text in a manner that was diverse, complex, and site specific. Medication data were typically unstructured and available only within free-text notes. More recently, computer-based prescribing has facilitated the entry of medication data in a format that is coded. National efforts are currently under way to harmonize both these data formats.[18]

Early in the past decade, lack of structured medication data presented an obstacle to the routine use of EMRs in studies addressing the genetic determinants underlying treatment outcomes. Therefore, initial progress in the field of pharmacogenetics was limited primarily to cohorts enrolled in treatment trials.[19] However, as clinical interest in this field grew, investigators began manually interrogating clinical practice-based data sets for drug-exposed cohorts of limited sample sizes.[20] Within this setting, the process of case ascertainment was labor intensive and expensive. Nonetheless, such studies were partly successful, particularly when they focused on predictors of toxicity (rather than efficacy) for drugs associated with clinically severe adverse drug reactions (ADRs) and a narrow therapeutic index.[21–23]

More recently, the electronic reconstruction of comprehensive medication histories has allowed the field of pharmacogenomics to gain considerable momentum within data sets that are derived from routine clinical practice.[9] Data derived from EMRs have proven to be highly accurate for quantifying disease phenotypes (onset and rate of progression) as well as treatment outcomes (efficacy and toxicity). Clinical diagnoses can be efficiently extracted from de-identified EMRs through the application of algorithms integrating diagnostic codes, clinical laboratory data, and medication histories, and these traits are robust in their ability to replicate known associations previously characterized in disease-specific research cohorts.[15] With the development of controlled vocabularies, their inclusion in

centralized terminology systems such as the Unified Medical Language System, and application of scalable information extraction strategies from the clinical narrative[24] (http://www.ohnlp.org; http://www.i2b2.org), investigators are now able to extend this work to the characterization of treatment outcomes using NLP software.[25]

Standard evaluation metrics for the performance of NLP systems are shown in Eqs 1–4:

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \quad (1)$$

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \quad (2)$$

$$F\text{-score} = \frac{2(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (3)$$

$$\text{Accuracy} = \frac{\text{True positives} + \text{True negatives}}{\text{True positives} + \text{False positives} + \text{False negatives} + \text{True negatives}} \quad (4)$$

Detailed medication information including strength of dose, route of administration, and frequency of administration can now be extracted from clinical text data. Jagannathan *et al.*[26] recently compared four commercial NLP engines and reported a high *F*-score of 93.2% for capturing drug names, but lower *F*-scores of 85.3, 80.3, and 48.3% for retrieving data on strength of dose, route of administration, and frequency of administration. Newer approaches can increase *F*-scores to >90% for strength, route, and frequency.[9] A recent NLP challenge in the Integrating Biology and the Bedside (i2b2) initiative focuses specifically on the extraction of drug signatures from clinical free-text documents (https://www.i2b2.org/NLP/Medication/).

CURRENT EFFORTS

Clearly, for pharmacogenetic and pharmacogenomic association studies to be successful within the context of EMRs, accurate drug-exposure histories need to be efficiently extracted and linked to treatment outcomes that are carefully defined. Methods used to quantify the toxicity of a given drug are typically quite different from those used to quantify its clinical efficacy.[27] As mentioned previously, many of the early findings in this field were related to toxicity.[21–23]

EMRs and drug toxicity

Data obtained from routine medical practice are uniquely suited for the investigation of genetic determinants underlying ADRs, whereas, in comparison, data obtained from clinical trials may be less useful for this task.[28] This discrepancy is due, in part, to the fact that clinical trials often have a "run-in" period, during which potential study subjects who show signs of drug intolerance are excluded before randomization. Patients with relevant comorbidities may also be actively excluded from clinical trials. This is not so in practice-based data sets. As a result, the prevalence of ADRs within the community is often higher (and more variable) than that observed within randomized trials.

Consider the example of lipid-lowering therapy. HMG-CoA reductase inhibitors (statins) are the most commonly prescribed class of medications in the United States, and large multicenter trials have demonstrated unequivocally that these drugs reduce the risk of cardiovascular events in patients at risk. Although statin-related ADRs occur very infrequently in the context of monotherapy, the ADR event rate increases within the context of interacting medications.[29,30] Many of these drug–drug interactions reflect alterations in pharmacokinetic processes.

We have also observed that genetic variability in pharmacokinetic processes can contribute to the severity of statin-related ADRs.[20] However, patient-to-patient variability in phase I metabolism represents only a single component within any given patient's capacity for drug disposition. The entire process is better understood when each component is considered within the context of absorption, distribution, metabolism, and elimination.[31] For example, many statins undergo additional modification through phase II conjugation by enzymes within the UDP-glucuronosyltransferase (UGT) family.[32] UGT1A1 and UGT1A3 are both capable of converting atorvastatin acid to a lactone derivative, and perturbations in atorvastatin kinetics previously attributed to UGT1A1*28 may in fact be due to genetic variability in UGT1A3 (UGT1A3 haplotypes are in allelic association with UGT1A1*28).[33] Membrane transporters also markedly influence statin disposition. The SEARCH Collaborative Group recently reported that simvastatin-induced muscle toxicity is associated with genetic variability in statin uptake.[34] Other data suggest that variability in efflux may influence risk.[35] Efforts are under way to replicate these findings using EMRs.[30]

The ability to resolve genetic determinants of drug toxicity depends on several factors.[22] Unless the toxicity end point is rigorously defined, such studies are subject to misclassification bias. Two important properties must be considered: the clinical severity of the ADR and the therapeutic index of the drug. The therapeutic index reflects the ratio of the dose known to cause half of the maximal toxicity to the dose known to deliver half of the maximal efficacy (TD_{50}/ED_{50}).[31] Drugs with a wider therapeutic

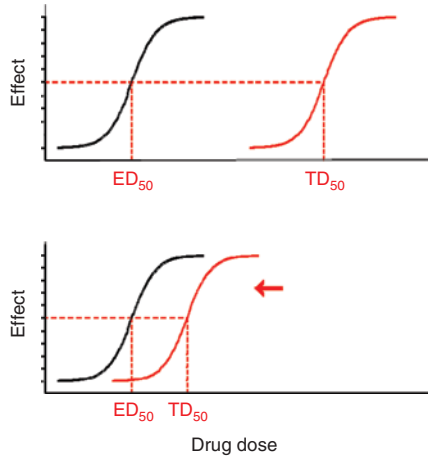

Figure 2 Quantifying drug toxicity. Therapeutic index (TI) = TD_{50}/ED_{50}. ED_{50} = dose of a drug observed to yield half-maximal efficacy. TD_{50} = dose of a drug observed to yield half-maximal toxicity.

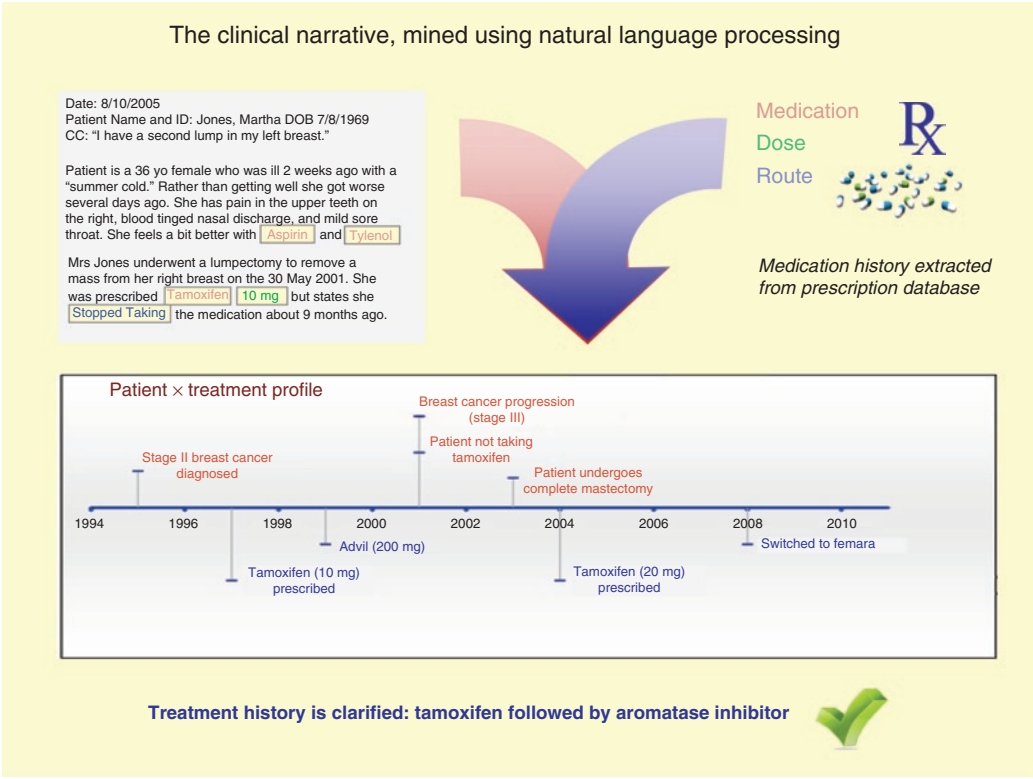

Figure 3 Structured and unstructured data generate high-quality phenotypes. Upper left: recent advances in natural language processing (NLP) allow extremely accurate reconstruction of comprehensive medication histories. Upper right: structured medication data generated by computerized provider order entry software (e.g., name–value pairs, such as "medication = tamoxifen") can be easier to collate and analyze. However, structured data must be normalized across diverse systems of care. The National Library of Medicine (NLM) has developed a terminology called RxNorm (http://www.nlm.nih.gov/research/umls/rxnorm), linking drug names (dose, ingredient, and formulation) to drug vocabularies commonly used in pharmacy management systems (e.g., First Databank, Micromedex, MediSpan, Gold Standard Alchemy, and Multum). Bottom: structured and unstructured data can be merged to yield high-quality drug-exposure phenotypes that facilitate pharmacogenomic studies using EMRs. (The hypothetical patient and clinical note featured above are completely fictitious.)

index (e.g., statins) tend to have ADRs that are less susceptible to genetic (or environmental) perturbations in kinetic processes. This may explain, in part, why statin-related ADRs occur so rarely within the context of monotherapy.[22] Conversely, ADRs tend to occur more frequently with drugs that have a narrower TI (e.g., anticoagulants and antineoplastics). This principle is illustrated in **Figure 2**.

To appreciate the importance of therapeutic index, one might consider the thromboembolic complications related to the use of tamoxifen, an antiestrogen used to treat breast cancer. Worldwide, it is the most common endocrine therapy for estrogen receptor–positive breast cancer. Although aromatase inhibitors are also commonly prescribed for this indication, ~1.5 million tamoxifen prescriptions are still written in the United States each year. Tamoxifen reduces the risk of breast cancer recurrence by ~50%, but there is considerable variability in its efficacy and in its toxicity profile.

Deep venous thromboses are rare but serious side effects linked to the use of tamoxifen. Because they are rare, there may not be enough events in typical clinical trials to conduct adequately powered genotype–phenotype association studies. Therefore, in order to study this association, EMRs have recently been used to identify sufficient numbers of patients with breast cancer who experience deep venous thromboses while on tamoxifen. Through investigations using banked DNA from ADR case patients and frequency-matched tamoxifen-exposed controls, it was found that genetic variants in the estrogen receptor were associated with the occurrence of deep venous thromboses.[36] Studies are now under way to confirm these results, using additional EMRs linked to biobanks participating in the National Institutes of Health–funded eMERGE network (https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page).

EMRs and drug efficacy

Tamoxifen is a moderately potent antiestrogen; however, it is metabolized into many metabolites that are known to be more potent than the parent compound. The most abundant of these high-potency metabolites is endoxifen (4-OH-*N*-desmethyl-tamoxifen). This active metabolite is formed primarily by cytochrome P450 2D6 (CYP2D6). As expected, patients have reduced concentrations of endoxifen in plasma if they have reduced CYP2D6 activity because of either the presence of genetic polymorphisms in CYP2D6 or the concurrent use of medications that inhibit CYP2D6 activity. Some association studies have shown a relationship between CYP2D6 activity and tamoxifen treatment outcome. In this context, EMRs have been used in recent efforts to determine whether the concurrent use

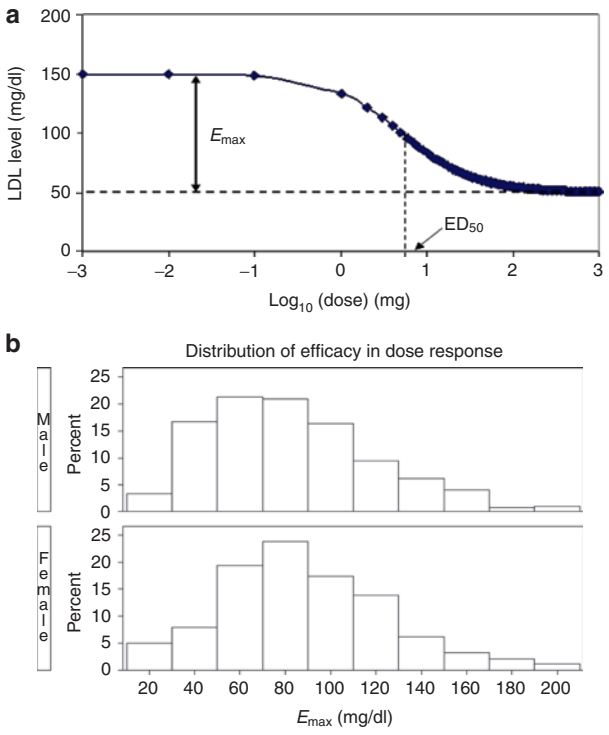

Figure 4 Quantifying drug efficacy within large populations, using electronic medical records (EMRs). (**a**) Dose–response data for atorvastatin. LDL cholesterol plotted by dose. (**b**) Gender-stratified distribution for E_{max} within an EMR biobank.

of CYP2D6-inhibiting drugs (e.g., antidepressants) by patients taking tamoxifen is associated with an increase in the recurrence rates of breast cancer. These studies have drawn from breast cancer health records from Denmark,[37] the Netherlands,[38] Canada,[39] and other countries. Another study utilized records from the Medco, Inc., pharmacy benefits manager in the United States.[40] These studies have shown mixed results. In some of the studies, but not in others, CYP2D6 inhibitors were reported to be associated with recurrence rates of breast cancer. The discrepancy in results may be due to differences in the populations studied, sample sizes, and/or the methods used to mine the databases. This highlights the need for additional work focused on understanding the optimal and consistent use of EMRs.

Within the National Institutes of Health–funded Pharmacogenomics Research Network (PGRN) (http://www.PharmGKB.org), in biobanks linked to EMRs, there are several ongoing efforts to construct and standardize comprehensive treatment histories of patients exposed to tamoxifen. Our group has recently applied automated NLP-based extraction and information merging within the EMR to quantify each patient's breast cancer treatment strategy within several Pharmacogenomics Research Network nodes. Sensitivity and specificity were 90.27–99.73% (positive and negative predictive values were 80.00–99.93%) as compared with a manually abstracted data set.

Figure 3 illustrates how the combination of prescription database querying and NLP techniques applied to the clinical narrative can be merged to assemble an automated treatment classification. Similar efforts are under way in other systems of care[41] and for other drug–gene interactions.[42] McAlpine and colleagues have effectively scanned EMR data to quantify drug exposure (including dose) as part of an investigation to determine the strength of association between genetic variability in CYP2D6 and therapeutic response to antidepressants.

By altering drug concentration levels in the circulation (or within a tissue microenvironment), many genetic predictors of toxicity are also known to alter drug efficacy (e.g., variable drug oxidation by cytochromes P450). Genetic variability in phase I metabolism of some statins not only influences the severity of ADRs (as mentioned earlier) but also affects the degree to which these drugs decrease low-density lipoprotein cholesterol levels.[43] Given that the impact of size appears to be relatively small (<10 mg/dl difference in cholesterol lowering per copy of the minor allele), the clinical significance of these findings remains uncertain.[43]

Variability in pharmacodynamic candidate genes can have similar effects. Variants in HMG-CoA reductase clearly influence the lipid-lowering efficacy of statins.[44] Large, multicenter studies are now under way to identify additional genes that influence the efficacy of statins; this research is being carried out primarily through the retrospective genotyping of archived biological materials obtained during prior randomized clinical trials.[11] However, restricting these efforts to treatment trials would produce only very limited information because most efficacy data obtained during trials are limited to a single dose. Complete characterization of drug response requires a consideration of potency (ED_{50}) as well as of maximal efficacy. These properties can be determined only in patients who are exposed to multiple drug doses and may best be derived from real-life, clinical practice–based data.

Without a complete characterization of efficacy in subjects exposed to drugs across a wide range of doses, clinical phenotypes cannot be fully characterized. To begin identifying genetic markers associated with the low-density lipoprotein cholesterol–lowering effect of atorvastatin across multiple doses, we previously interrogated EMR data from a large population-based biobank.[27] NLP software was used to generate comprehensive retrospective drug-exposure histories for the entire database (n = 20,000). For statin exposure, these algorithms were 100% sensitive and 96% specific, with an initial positive predictive value of 87%. Through manual chart abstraction, these algorithms were optimized using programming that corrected for dosing discrepancies attributed to pill splitting; the final positive predictive value was 95%. Full dose–response relationships were then constructed for all biobank participants exposed to atorvastatin. The result was a nested cohort of 3,710 individuals for whom we subsequently derived rigorous phenotypic parameters for atorvastatin potency (ED_{50}) and maximal atorvastatin efficacy (E_{max}).[27] The distribution of these traits is shown in **Figure 4**.

The application of algorithms rigorously characterizing efficacy within EMR data is now being expanded to other classes of drugs across diverse systems of care. Through an initiative led by the PGRN (http://www.pharmgkb.org/contributors/pgrnResources/pgpop_profile.jsp), these approaches are being

explored for the accurate characterization of treatment outcomes within multiple nodes of the HMO Research Network. Network members include Harvard Pilgrim and Fallon Healthcare (in the northeastern United States); Kaiser Permanente Georgia (in the southeastern United States); HealthPartners, Henry Ford, and Marshfield Clinic (in the midwestern United States); and Group Health Cooperative, Lovelace Clinic, Kaiser Permanente Colorado, Kaiser Permanente Northwest, Kaiser Permanente Hawaii, Kaiser Permanente Southern California, and Kaiser Permanente Northern California (in the western United States). These institutions provide care to more than 10 million individuals.[8]

Cross-network consortia are currently being established, and more than 250,000 DNA samples are already in hand. Kaiser Permanente Northern California, in particular, has recently embarked on one of the most ambitious biobanking efforts ever undertaken. As biological materials are linked to additional EMRs in these and other systems of care, the resulting data will advance the field of pharmacogenomics in two distinct ways: first, by enabling assessment of the generalizability of previously identified drug response genes within the community, and second, by facilitating discovery of previously unrecognized determinants of drug outcome in the context of clinical covariates.

VISION FOR THE FUTURE

The knowledge base relating variable outcomes in drug therapy to genomic variations is rapidly expanding. As noted above, the construction and integration of biobanks linked to EMRs can follow a variety of designs. Biobanks that enroll subjects through the process of informed consent must comply with laws regarding privacy at the local, state, and federal levels.[10,12] Clinical data and biological materials that are de-identified in accordance with the provisions of 45 CFR 46 may be used for research with "nonhuman subjects."[15] As these and other EMR-based data sources are merged for large-scale pharmacogenomic studies, all such efforts must be subject to ongoing oversight by institutional review boards, internal and external ethics committees, community advisory boards, legal departments, and the Federal Office of Human Research Protection.

Data security and privacy are paramount. The Safe Harbor provision of the HIPAA Privacy Rule requires the removal of 18 personal identifiers (including demographic data) before any form of public disclosure. Additional layers of security are being developed and tested as the clinical and scientific communities manage increasingly complex genomic data sets.[45]

Decision support

As efforts to merge biobanks grow, there is increasing difficulty in envisioning how pharmacogenomic knowledge can be brought to the bedside in the absence of advanced informatics tools that would include examination of the levels of evidence as well as advice on how to manage individual subjects. The scientific and clinical communities have only begun to leverage the wide variety of phenotypic data available in EMRs to optimize studies of drug outcomes. Medication information from the structured (e.g., electronic prescribing systems) and free-text components (e.g., from clinical notes) of each EMR will need to be merged, as shown in **Figure 3**. Additional data sources will allow assessment of adherence to medication regimens. Pharmacy claim data have been standardized for many years and are widely used in health services research, especially for Pharmacy Benefit Management programs, Medicaid, and, recently, Part D of Medicare.

Data normalization strategies (using locally adopted formats and terminologies) need to be shared across large healthcare networks. The mapping of clinically relevant terms to community-vetted and -adopted ontologies will ensure semantic interoperability and ease of study replication. Extending the example developed in **Figure 3**, variations in the terms indicating exposure to tamoxifen can be mapped to the same RxNORM code (10324) so that investigators may avoid the creation of exhaustive lists of terms referring to the same phenotypic data element.

Normalization procedures such as these will enable the establishment of large patient registries, with unique biological and clinical characteristics, across institutions. Supported by National Institutes of Health funding, many medical centers are already applying NLP techniques for extracting information from the clinical narrative to investigate rare medication side effects prospectively. Although the need for NLP analysis can be reduced by an increase in structured data entry in EMRs, novel approaches to the characterization of free text will continue to be essential. For example, although structured applications in medical oncology now automatically encode the chemotherapy regimens ordered (and medications actually delivered to a patient), free-text interrogation is still needed to understand why divergence occurs, in some cases, between the medications ordered and those actually delivered.

The rapid advancements in electronic phenotyping approaches, described earlier, have occurred in parallel with equally robust progress in the area of genotyping technologies. Genome-wide SNP scanning arrays have become increasingly dense (currently containing more than 1 million SNPS) and increasingly cost effective over the past decade. As a result, more genetic information is available at lower cost. Exome scanning (i.e., a focused approach that sequences all exons for all genes) is being conducted in cohorts of increasing sample sizes, and it seems inevitable that entire genomes will soon be included in each patient's EMR.[1,2]

Ultimately, the clinical utility of this phenotypic and genotypic information will depend on biomedical informatics application platforms that provide efficient real-time decision support at the point of care. Furthermore, the expansion of knowledge generated by ongoing genotype–phenotype association studies will make it imperative that decision-support platforms be flexible enough to incorporate future knowledge. Decision-support software must allow reinterpretation of clinical genetic data in the light of information that may emerge from future discoveries.[46] There is also a considerable need to educate and train health-care professionals in the use of these data. A recent survey carried out by the

American Medical Association in collaboration with Medco Health Solutions, Inc., found that, although 98% of provider participants agreed about the utility of genetic testing in drug therapy, only 10% actually felt that they had been adequately informed about the process. Only 26% had received some form of formalized training (https://www.medcoresearch.com/community/pharmacogenomics/physicansurvey). This stands in strong contrast to interview data that clearly indicate that most patients expect health-care professionals to explain the clinical utility of pharmacogenetic tests.[47]

Clinical implementation

To ease the transition of this information into routine clinical practice, Gardner *et al.* have proposed the addition of genetic information to an existing drug interaction database so as to further enhance clinical decision support and optimize patient safety.[48] This approach is appealing from a clinical viewpoint. Relevant genetic data can be formatted as an Extensible Markup Language document, and the additional information can be added using tags and implemented through existing software. Such software is already in place at many large medical centers. The real challenge will be in deciding when and how to use this information. Potential application paradigms include (i) gene-based drug selection, (ii) gene-based drug dosing, (iii) medication reconciliation, (iv) "push" phenotyping, and (v) pharmacovigilance.

Warfarin provides a good example of the use of genetic information in determining drug dosing. Clearly an individual's maintenance dose of warfarin is strongly influenced by variability in the pharmacodynamics and pharmacokinetics of the drug induced by the presence of specific candidate genes, and this relationship can be utilized to inform the process of gene-based drug dosing. Although similar strategies are being developed for managing the effect of pharmacokinetics-altering candidate genes on the efficacy of clopidogrel, gene-based drug selection (e.g., prescribing another thienopyridine in place of clopidogrel) is a suitable alternative. It is important to note, however, that these strategies need not be limited to the period of drug initiation. Both can be applied, in the context of a variety of drugs, during the process of medication reconciliation, particularly when patients move from a system of care without an EMR to one with an EMR.[49]

Furthermore, even in the absence of genetic information, EMRs can facilitate gene-based drug dosing by flagging in real-time the data of patients who appear to be developing an ADR and by prompting providers to consider genotyping such patients in an effort to optimize risk assessment. EMRs can also be used at the population level to identify trends in the development of new ADRs; that is, EMRs may enhance pharmacovigilance, especially during the postmarketing period. For instance, there are now compelling data to support the claim that cardiovascular ADRs related to the use of highly potent COX-2 inhibitors might have been identified earlier using such an EMR-based approach.[50]

In 2007, the US Congress passed the Food and Drug Administration Amendments Act, directing the agency to increase the postmarket monitoring of drug safety. In order to accomplish this, the Food and Drug Administration established the "Sentinel Initiative" with the goal of detecting drug safety–related signals earlier and more accurately. This is now being accomplished through real-time monitoring of EMRs and medical claims databases. The monitoring of multiple databases throughout the health-care system requires the sophisticated integration of diverse EMR databases. Although many cross-institutional networks have already been constructed to facilitate this process (e.g., the PGRN, the eMERGE network, and the HMO Research Network), it is likely that new funding initiatives and patient advocacy groups will drive the implementation of even larger consortia. Such efforts will generate the data needed to define the role of EMRs in pharmacogenomics and drug safety.

Outlook

Data on pharmacogenetic associations of various effect sizes are streaming into the literature at an increasing rate. It is not possible for individual practitioners to keep track of these relationships without assistance from information technology systems. However, with electronic decision support, clinical practitioners now have the ability to utilize this information. Genetic data will likely soon be deposited preemptively into each patient's EMR, and robust biomedical informatics platforms will be positioned to interrogate this information during the process of clinical decision making in real time. It appears that EMRs have brought personalized medicine to our doorstep.

ACKNOWLEDGMENTS

This work was supported by the following NIH grants: R01DK080007, U01HL069757, U01HG004608, U01 HL65962, RC2GM092618, R01CA139246, NCI DCCPS supplement to U01 HG 04599, UO1GM61388, Breast SPORE CA 116201, U54LM008748, R01GM088076, U01GM061373, and UL1 RR024975.

CONFLICT OF INTEREST

The authors declared no conflict of interest.

1. Ashley, E.A. *et al*. Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).
2. Lifton, R.P. Individual genomes on the horizon. *N. Engl. J. Med.* **362**, 1235–1236 (2010).
3. DesRoches, C.M. *et al*. Electronic health records in ambulatory care—a national survey of physicians. *N. Engl. J. Med.* **359**, 50–60 (2008).
4. Jha, A.K. *et al*. Use of electronic health records in U.S. hospitals. *N. Engl. J. Med.* **360**, 1628–1638 (2009).
5. Shea, S. & Hripcsak, G. Accelerating the use of electronic health records in physician practices. *N. Engl. J. Med.* **362**, 192–195 (2010).
6. Pakhomov, S., Bjornsen, S., Hanson, P. & Smith, S. Quality performance measurement using the text of electronic medical records. *Med. Decis. Making* **28**, 462–470 (2008).
7. Denny, J.C., Miller, R.A., Waitman, L.R., Arrieta, M.A. & Peterson, J.F. Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. *Int. J. Med. Inform.* **78 (suppl. 1)**, S34–S42 (2009).
8. Chan, K.A. *et al*. The HMO Research Network. In *Pharmacoepidemiology* 4th edn., (ed. Strom, B.L.) 261–270 (Wiley & Sons, West Sussex, UK, 2007).
9. Xu, H., Stenner, S.P., Doan, S., Johnson, K.B., Waitman, L.R. & Denny, J.C. MedEx: a medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.* **17**, 19–24 (2010).
10. McCarty, C.A. & Wilke, R.A. Biobanking and pharmacogenomics. *Pharmacogenomics* **11**, 637–641 (2010).
11. Barber, M.J. *et al*. Genome-wide association of lipid-lowering response to statins in combined study populations. *PLoS ONE* **5**, e9763 (2010).

12. McCarty, C.A. *et al*. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *Personalized Med* **2**, 49–79 (2005).
13. Roden, D.M. *et al*. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008).
14. Ormond, K.E., Cirino, A.L., Helenowski, I.B., Chisholm, R.L. & Wolf, W.A. Assessing the understanding of biobank participants. *Am. J. Med. Genet. A* **149A**, 188–198 (2009).
15. Ritchie, M.D. *et al*. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* **86**, 560–572 (2010).
16. McDonald, C.J. Protocol-based computer reminders, the quality of care and the non-perfectability of man. *N. Engl. J. Med.* **295**, 1351–1355 (1976).
17. McDonald, C.J., Murray, R., Jeris, D., Bhargava, B., Seeger, J. & Blevins, L. A computer-based record and clinical monitoring system for ambulatory care. *Am. J. Public Health* **67**, 240–245 (1977).
18. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C. & Hurdle, J.F. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb. Med. Inform.* 128–144 (2008).
19. Evans, W.E., Relling, M.V., Rodman, J.H., Crom, W.R., Boyett, J.M. & Pui, C.H. Conventional compared with individualized chemotherapy for childhood acute lymphoblastic leukemia. *N. Engl. J. Med.* **338**, 499–505 (1998).
20. Wilke, R.A., Moore, J.H. & Burmester, J.K. Relative impact of CYP3A genotype and concomitant medication on the severity of atorvastatin-induced muscle damage. *Pharmacogenet. Genomics* **15**, 415–421 (2005).
21. Evans, W.E. & Relling, M.V. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**, 487–491 (1999).
22. Wilke, R.A. *et al*. Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. *Nat. Rev. Drug Discov.* **6**, 904–916 (2007).
23. Klein, T.E. *et al*.; International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Med.* **360**, 753–764 (2009).
24. Friedman, C. A broad-coverage natural language processing system. *Proc. AMIA Symp.* 270–274 (2000).
25. Chhieng, D., Day, T., Gordon, G. & Hicks, J. Use of natural language programming to extract medication from unstructured electronic medical records. *AMIA Annu. Symp. Proc.* 908 (2007).
26. Jagannathan, V. *et al*. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. *Int. J. Med. Inform.* **78**, 284–291 (2009).
27. Wilke, R.A., Berg, R.L., Linneman, J.G., Zhao, C., McCarty, C.A. & Krauss, R.M. Characterization of low-density lipoprotein cholesterol-lowering efficacy for atorvastatin in a population-based DNA biorepository. *Basic Clin. Pharmacol. Toxicol.* **103**, 354–359 (2008).
28. Vandenbroucke, J.P. & Psaty, B.M. Benefits and risks of drug treatments: how to combine the best evidence on benefits with the best data about adverse effects. *JAMA* **300**, 2417–2419 (2008).
29. McClure, D.L., Valuck, R.J., Glanz, M., Murphy, J.R. & Hokanson, J.E. Statin and statin-fibrate use was significantly associated with increased myositis risk in a managed care population. *J. Clin. Epidemiol.* **60**, 812–818 (2007).
30. Mareedu, R.K. *et al*. Use of an electronic medical record to characterize cases of intermediate statin-induced muscle toxicity. *Prev. Cardiol.* **12**, 88–94 (2009).
31. Wilke, R.A., Reif, D.M. & Moore, J.H. Combinatorial pharmacogenetics. *Nat. Rev. Drug Discov.* **4**, 911–918 (2005).
32. Prueksaritanont, T., Tang, C., Qiu, Y., Mu, L., Subramanian, R. & Lin, J.H. Effects of fibrates on metabolism of statins in human hepatocytes. *Drug Metab. Dispos.* **30**, 1280–1287 (2002).
33. Riedmaier, S. *et al*. UDP-glucuronosyltransferase (UGT) polymorphisms affect atorvastatin lactonization *in vitro* and *in vivo*. *Clin. Pharmacol. Ther.* **87**, 65–73 (2010).
34. Link, E. *et al*.; SEARCH Collaborative Group. SLCO1B1 variants and statin-induced myopathy–a genomewide study. *N. Engl. J. Med.* **359**, 789–799 (2008).
35. Keskitalo, J.E., Zolk, O., Fromm, M.F., Kurkinen, K.J., Neuvonen, P.J. & Niemi, M. ABCG2 polymorphism markedly affects the pharmacokinetics of atorvastatin and rosuvastatin. *Clin. Pharmacol. Ther.* **86**, 197–203 (2009).
36. Onitilo, A.A. *et al*. Estrogen receptor genotype is associated with risk of venous thromboembolism during tamoxifen therapy. *Breast Cancer Res. Treat.* **115**, 643–650 (2009).
37. Lash, T.L. *et al*. Breast cancer recurrence risk related to concurrent use of SSRI antidepressants and tamoxifen. *Acta Oncol.* **49**, 305–312 (2010).
38. Dezentjé, V.O. *et al*. Effect of concomitant CYP2D6 inhibitor use and tamoxifen adherence on breast cancer recurrence in early-stage breast cancer. *J. Clin. Oncol.* **28**, 2423–2429 (2010).
39. Kelly, C.M. *et al*. Selective serotonin reuptake inhibitors and breast cancer mortality in women receiving tamoxifen: a population based cohort study. *BMJ* **340**, c693 (2010).
40. Aubert, R.E. *et al*. Risk of breast cancer recurrence in women initiating tamoxifen with CYP2D6 inhibitors. *J Clin Oncol* **27**, 9s abstr CRA508 (2009).
41. Liao, K.P. *et al*. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res. (Hoboken).* **62**, 1120–1127 (2010).
42. McAlpine, D.E., O'Kane, D.J., Black, J.L. & Mrazek, D.A. Cytochrome P450 2D6 genotype variation and venlafaxine dosage. *Mayo Clin. Proc.* **82**, 1065–1068 (2007).
43. Kivistö, K.T. *et al*. Lipid-lowering response to statins is affected by CYP3A5 polymorphism. *Pharmacogenetics* **14**, 523–525 (2004).
44. Krauss, R.M. *et al*. Variation in the 3-hydroxyl-3-methylglutaryl coenzyme a reductase gene is associated with racial differences in low-density lipoprotein cholesterol response to simvastatin treatment. *Circulation* **117**, 1537–1544 (2008).
45. Loukides, G., Gkoulalas-Divanis, A. & Malin, B. Anonymization of electronic medical records for validating genome-wide association studies. *Proc. Natl. Acad. Sci. USA* **107**, 7898–7903 (2010).
46. Mitchell, D.R. & Mitchell, J.A. Status of clinical gene sequencing data reporting and associated risks for information loss. *J. Biomed. Inform.* **40**, 47–54 (2007).
47. Fargher, E.A. *et al*. Patients' and healthcare professionals' views on pharmacogenetic testing and its future delivery in the NHS. *Pharmacogenomics* **8**, 1511–1519 (2007).
48. Gardner, D. Using genomics to help predict drug interactions. *J. Biomed. Inform.* **37**, 139–146 (2004).
49. Bassi, J., Lau, F. & Bardal, S. Use of information technology in medication reconciliation: a scoping review. *Ann. Pharmacother.* **44**, 885–897 (2010).
50. Brownstein, J.S., Sordo, M., Kohane, I.S. & Mandl, K.D. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. *PLoS ONE* **2**, e840 (2007).