Electronic health records: Implications for drug discovery

Lixia Yao[1], Yiye Zhang[2], Yong Li[1], Philippe Sanseau[3] and Pankaj Agarwal[1]

[1] Computational Biology, GlaxoSmithKline R&D, King of Prussia, PA 19406, USA
[2] Department of Biostatistics, Columbia University, New York, NY 10032, USA
[3] Computational Biology, GlaxoSmithKline R&D, Stevenage SG1 2NY, UK

Electronic health records (EHRs) have increased in popularity in many countries. Pushed by legal mandates, EHR systems have seen substantial progress recently, including increasing adoption of standards, improved medical vocabularies and enhancements in technical infrastructure for data sharing across healthcare providers. Although the progress is directly beneficial to patient care in a hospital or clinical setting, it can also aid drug discovery. In this article, we review three specific applications of EHRs in a drug discovery context: finding novel relationships between diseases, re-evaluating drug usage and discovering phenotype–genotype associations. We believe that in the near future EHR systems and related databases will impact significantly how we discover and develop safe and efficacious medicines.

Introduction

In the 1960s researchers at academic medical centers and government healthcare departments embarked on transforming article medical records into electronic medical records (EMRs) [1]. EMRs were largely conceived as a collection of records kept at a single healthcare organization. They have, however, expanded into longitudinal electronic records of patient health information generated during one or more encounters in any care delivery setting and these longitudinal records are now known as electronic health records (EHRs). EHRs often contain patient demographics, progress notes, problem notes, medication information, vital signs, past medical history, immunizations, laboratory data and radiology reports (Fig. 1). With their promise of reducing healthcare costs, improving quality of care and promoting evidence-based medicine, EHRs have been adopted in a growing number of countries, including but not limited to Australia, Canada, UK, The Netherlands, New Zealand and the USA [2]. In the USA the adoption rate of basic EHR systems has grown to cover 44% of all physicians [3] and will grow even faster with governmental policies that require ‘meaningful use’ of EHRs [4]. Other countries, such as Estonia, Denmark and Singapore, have been implementing nationwide EHR systems [5,6]. The potential for growth in the adoption and utilization of EHRs is large given active research and progress in the areas of standards, terminology, security and confidentiality, natural language processing and telemedicine [7–12]. In this article, we focus on emerging and potential applications of EHRs in the context of drug discovery.

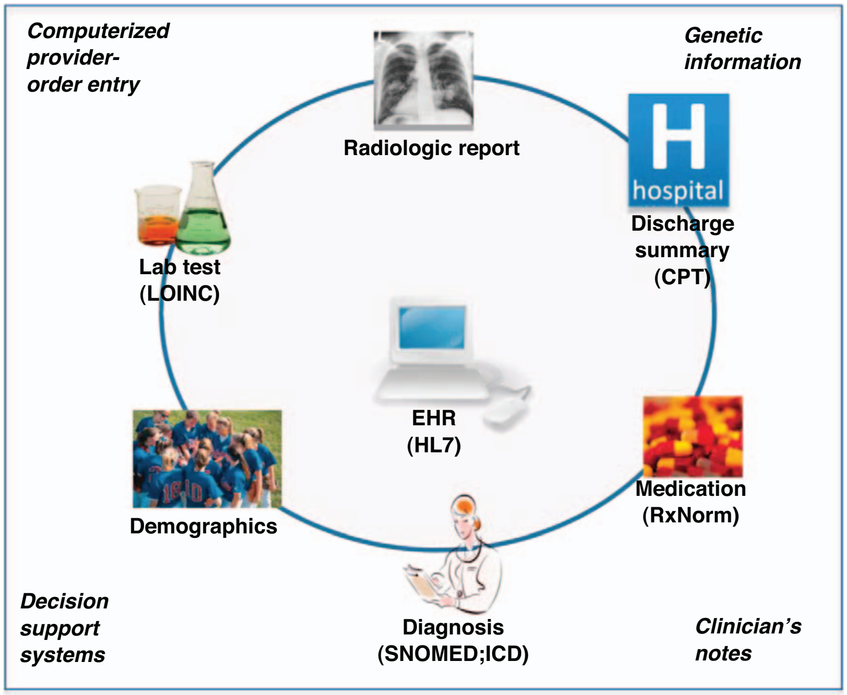
EHR databases (Table 1) are routinely used in the pharmaceutical industry for market research, pharmacovigilance, clinical marker validation and drug safety evaluation. However, because EHRs provide observational data for a large population over long periods of time, it is possible to utilize them for a deeper understanding of how drugs affect patients through changes in diagnosis, disease progression and laboratory measurements. There is an increasing appreciation of the benefits of mining EHR information for drug discovery, such as the detection of disease relationships, drug repositioning and genotype–phenotype association discovery. All these translational research efforts have direct and important implications for early drug discovery and could ultimately lead to safer and more-efficacious medicines (Fig. 2).

Using EHRs to identify novel disease relationships

In concert with molecular and genetic data, population-based disease relationships or comorbidities can be used to elucidate the mechanisms of complex diseases and to discover new treatments. Relationships within a disease network can highlight

Corresponding author:. Agarwal, P. (pankaj.agarwal@gsk.com)

1359-6446/06/$ - see front matter © 2011 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2011.05.009



Drug Discovery Today

FIGURE 1

Elements and standards for EHR systems. A basic EHR system contains clinical information such as demographic characteristics of patients, diagnoses, medication information, hospitalization and discharge summaries, radiologic reports and laboratory tests. More-advanced EHR systems can also include clinician's notes, and functionalities such as computerized provider order entry and decision support systems, such as clinical guidelines, clinical reminders, drug-allergy alerts, drug–drug interaction alerts and drug–dose support. The commonly used standard and terminologies used in EHR systems are listed in brackets. HL7, or health level 7, is the global authority on standards for interoperability of health information technology.

Abbreviations: SNOMED, systematized nomenclature of medicine; ICD, international statistical classification of diseases and related health problems; RxNorm, a standardized nomenclature for clinical drugs and drug delivery devices produced by the National Library of Medicine; CPT, current procedural terminology; LOINC, logical observation identifiers names and codes.

TABLE 1

A selection of EHR databases

	Description	Reference/Link
Generic EHR database		
Clinical Data Services by GE Healthcare	16.7 million patients from the USA in the GE Centricity EMR database	https://www2.gehealthcare.com/portal/site/usen/menuitem.b399d8492e44a6765c09cbd58c829330/?vgnextoid=ae0f4fb9efff5210VgnVCM100000382b3903RCRD&vgnextfmt=default
German Mediplus @ IMS Health	12 million patients from primary practice in Germany	Ref. [49]
HealthConnect @ Kaiser Permanente	8.6 million patients covered by Kaiser Permanente's health plans	http://xnet.kp.org/newscenter/aboutkp/healthconnect/index.html
General Practice Research Database @ the Medicines Control Agency, UK	5 million active patients of research standard in the UK	http://www.gprd.com/home/
National Patient Care Database @ Veterans Health Administration (VHA)	Inpatient and outpatient services provided to 4 million VHA healthcare users in the USA	Ref. [50]
Clinical Data Warehouse @ Stanford University Medical Center (SUMC)	1.53 million pediatric and adult patients from 1994 to now at SUMC, including demographics, clinical encounters, diagnoses, clinical procedures, laboratory test results and inpatient pharmacy orders	http://clinicalinformatics.stanford.edu/projects/cdw.html
Clinical Data Repositories @ University of Virginia Health System	1 million patients over 15 years at University of Virginia	Ref. [51]
EHR database linked to DNA sample		
ORBIT @ Aurora Health Care	A library of blood samples linked to patient data for the purpose of testing new technologies	http://www.aurorahealthcare.org/services/orbit/bg.html
BioVU @ Vanderbilt University	80,000 medical records linked to DNA samples	Ref. [43]
MyCode @ Geisinger Health System	60,000 blood samples from over 23,000 Geisinger patients	http://www.geisinger.org/research/centers_departments/genomics/mycode/mycode.html

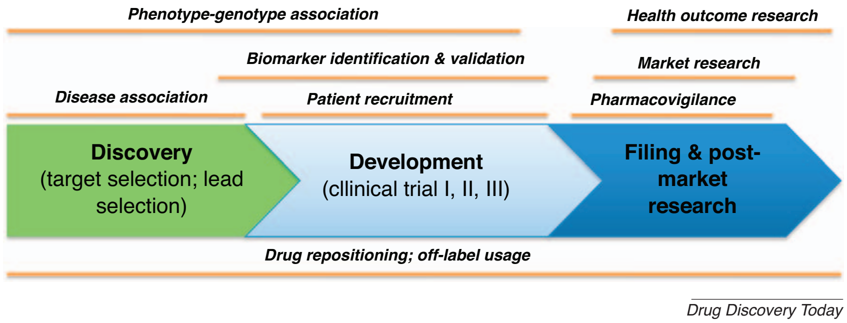

Phenotype-genotype association
Health outcome research
Biomarker identification & validation
Market research
Disease association
Patient recruitment
Pharmacovigilance
Discovery (target selection; lead selection)
Development (cllinical trial I, II, III)
Filing & post-market research
Drug repositioning; off-label usage
Drug Discovery Today

FIGURE 2

Various applications of EHR in drug discovery and development and the stages of the drug discovery and development pipeline that can be impacted.

common pathophysiological mechanisms, which could then be investigated to provide new insights into the disease etiology and lead to identification of new drug targets. More importantly, when a compound is available to treat one disease that belongs to a disease cluster other diseases from that cluster might represent new disease indications for that compound. EHR databases are among the richest sources of data for systematically identifying disease relationships, directionality and courses of disease progression. A combination of data mining of EHR databases and network analysis has demonstrated potential.

Hidalgo *et al.* analyzed >30 million MedPAR Medicare records covering 32 million elderly US citizens hospitalized between 1990 and 1993 [13]. They built a disease co-morbidity network, where the nodes represent diseases labeled with ICD-9 (International Statistical Classification of Diseases and Related Health Problems) codes and the 'edges' represent the strength of the disease relationships. The edges were determined using relative risk and Pearson's correlation for binary variables. Unsurprisingly, their results showed that patients tend to develop diseases that are close in the disease network to those that they already have, and that a disease that is connected to many other diseases tends to have a worse prognosis in comparison with diseases that are less connected. This would suggest that the need for better therapies is greater for patients with a more 'connected' disease. The network also revealed directionality of disease progression (i.e. diseases such as hypertension and ischemic heart disease are known to provoke the development of more diseases than average). We believe that investigation into co-occurring diseases that share drugs in comparison with those that do not, together with molecular pathway information, could aid validation of the methodology and potentially highlight new drugs for repositioning.

Hanauer *et al.* [14] also looked at EHR data but, rather than mining coded fields, they parsed ~1.5 million clinical problem summary lists (in free text) from 327,000 patients at the University of Michigan and calculated an odds ratio and *P* value for each diagnosis association pair, using a tool called Molecular Concept Map. They validated common known disease relationships and also found novel relationships, some of which were subsequently validated in the literature. For example, they reported a novel relationship between granuloma annulare and osteoarthritis with an odds ratio of 4.3 and a P value of 1.1×10^{-4}. However, both diseases are treatable with the same medicine – niacin [15,16]. Their finding suggested that the two diseases might share a common underlying biological pathway with potentially more shared drug targets.

A probability model of disease progression, based solely on EHR data, was recently used to simulate disease progression and infer genetic overlaps between disease phenotypes by Rzhetsky *et al.* [17]. The authors first estimated the parameters for the model using 1.5 million patient records at a large medical center. Then, by assuming for each pair of 161 selected diseases that they are either uncorrelated (independent), negatively correlated (genetic overlap via competition) or positively correlated (genetic overlap by cooperation), they found a three-way positive relationship among autism, bipolar disorder and schizophrenia, and suggested that these diseases share significant genetic overlaps. This finding was not unexpected because these diseases share similar symptoms and can be treated by common drugs. Additionally, the study also revealed a competitive genetic overlap between bipolar disorder and female breast cancer. Although there is no direct evidence supporting biological plausibility of its pharmacologic effect, tamoxifen, a breast cancer drug, was recently shown to be effective in treating symptoms of bipolar disorder [18] and mania [19].

Data-mining approaches to identify novel disease relationships from EHR databases, such as the ones described above, are growing; however, they are subject to several limitations. A statistical association does not imply medical relevance or a causal relationship and so, to make further inference, one would require more-sophisticated and -powerful methods that can take into account, for example, the temporal information within EHR databases and dependence of other factors [20]. In addition, all diagnoses held in EHRs are entered by the clinician or support staff in various healthcare setups based on a particular coding terminology. However, coded diagnoses are not used consistently and are often inaccurate or incomplete [21]. Further, internal hierarchical relationships between disease terms can be inconsistent, and semantic granularity needs to be reviewed manually before nonhypothesis-driven data-mining exploration. For example, ICD-9 has more than ten codes devoted to acute liver injury diagnosis.

Applying EHRs for drug usage re-evaluation

EHR databases contain drug prescription information and thus form a rich resource for understanding the interactions of drugs with patients. Mining drug prescription data, diagnosis/symptom narratives and demographic data in large EHR databases can enable the systematic identification of drug adverse events (pharmacovigilance), drug off-label use, new indications (also called drug repositioning or drug repurposing), new combination therapies and drug–drug interactions. Many of these events can be more difficult to observe in smaller populations over shorter periods of time in clinical trials.

Pharmacovigilance arose from the analysis of data from various spontaneous reporting systems enforced by regulatory agencies in different countries [22]. Pharmacovigilance using EHR databases has the advantage of providing sample size and denominators as opposed to the spontaneous reporting systems [21,23–27]. However, because pharmacovigilance is an active, well-established research area in epidemiology, we will not discuss it further in this article. Interested readers are referred to recent reviews by Johansson *et al.* [28] and Trifirò *et al.* [29].

Mining EHR data using methodologies similar to those in pharmacovigilance can also be used to identify 'desirable' side effects, or novel indications, which is the essential goal of drug repositioning [30]. This activity has gained momentum in the industry over the past few years and has become an important strategy for many pharmaceutical companies. One significant advantage of drug repositioning is in drug safety. The risk of failure because of poor safety is smaller for late-stage or marketed drugs than for early-stage compounds. This is primarily because late-stage assets and marketed drugs have already passed a significant number of toxicity tests in animal models, and there is also a better understanding of their side effects in patients. In addition, the time and associated costs required for early drug discovery are bypassed. Despite some potential intellectual property issues, drug repositioning carries the promise of significant societal benefits and continues to merit vigorous investigation.

Mining medication data in EHR databases for systematic screening of off-label usage can provide an insight into overlaps between disease biology and represents a pragmatic approach for generating new drug repositioning hypotheses, which need to be tested in clinical trials. The FDA defines off-label use of drugs as 'the use of a prescription drug for an indication, in a dosage form or dose regimen for a particular population in a way not stated in the approved labeling' [31]. In the USA, and many other countries, physicians can legally prescribe drugs, including many controlled substances, off-label. In fact, off-label use of medications appears to be common. More than 70% of patients in a recent study had off-label prescriptions [31]. It is, therefore, not surprising that 57% of new drug usages are found by clinicians through field discovery [32].

A retrospective study of a clinician's narrative notes for patients receiving tamoxifen in 2008, from a 650,000-member health maintenance organization (HMO) in Israel, found that the drug was used off-label in 5.8% of patients for unapproved indications including female infertility, ovarian cancer and prostate neoplasm [33]. A similar study at the Jefferson Headache Center in Pennsylvania found that the anticonvulsant topiramate has been used off-label for idiopathic intracranial hypertension [34]. The analysis was based on patient diagnoses coded in ICD-9 and recorded in free-text narratives contained in their in-house EHR database, and is consistent with a recent study by Celebisoy *et al.* [35]. Moreover, in a separate study of 75,389 computerized Medicaid administrative claim files, 52% of 1660 topiramate utilizations were found to be used off-label [31].

Notably, these projects are not based solely on data-mining techniques – most have a predefined biological or clinical hypothesis before statistical analysis begins and involve a significant amount of manual work. A key limitation for automatically discovering off-label use and new indications for known drugs from EHR data is that many existing EHR systems do not have fully coded diagnoses. Valuable information, such as the description of the medication outcome and temporal information, is often stored by clinicians in free-text narrative notes [36,37]. For coded elements the data quality is often not satisfactory [38]. Domain experts such as pharmacists and physicians are usually hired to examine the free-text section to screen out patients who do not meet the inclusion criteria, and to check manually the accuracy of the code assignments in ICD-9 or other terminologies for each diagnosis corresponding to the specific drug. Thus, improvement in current terminologies for diagnosis and drugs, and advanced natural language processing and text-mining technologies, are urgently needed to study systematically off-label use and drug repositioning in a large-scale automated fashion. In addition, based on our own experience and report from the Observational Medical Outcomes Partnership (OMOP) [39], the pharmacy or medication data in many EHR databases are presented in the form of a patient ID, date and event (prescription or refill) with incomplete dosage information. Such a data design makes it difficult to infer how long a patient has been on a given drug and to carry out temporal analysis.

Despite these challenges in methodologies and techniques, mining EHR databases for novel drug usage and possibly new synergistic combinations is a promising direction for future research. For instance, large clinical trials of combination therapies can be expensive because of multiple arms, and we hope EHR-based research can assist in identifying the most promising combination therapies.

Finding genotype–phenotype associations from EHRs

Genome-wide association studies (GWAS) examine genetic variations in thousands of individuals to investigate how these variations are associated with specific complex traits or phenotypes such as diseases. Collecting a well-characterized patient cohort of this size is often a rate-limiting step for many studies. Because EHR databases provide detailed patient demographics, longitudinal diagnoses, prescription history and laboratory data, cohorts for genetic studies have been identified using these resources [40,41]. Further, because EHR systems incorporate genetic data meeting confidentiality and patient de-identification requirements, they could even help to identify the set of clinical phenotypes associated with a given genotype (reverse GWAS), or perform a systematic scan to find more genotype–phenotype associations. An NIH-funded national consortium – eMERGE (electronic medical records and genomics) – has been created to combine patient DNA samples with EHR data for large-scale high-throughput genetic research [42].

Vanderbilt University's DNA databank, BioVU, is leading the way in the genetic associations in the EHR field. Researchers collect patient DNA samples from discarded blood samples from routine clinical tests. This resource is then linked to a de-identified EHR database at Vanderbilt University Medical Center. By March 2010 the resource included 80,635 DNA samples, with an accrual rate of $\sim$500–700 samples per week [43]. A proof of concept study recently derived case and control groups for five different diseases using an algorithm that groups billing codes, patient encounters and laboratory data, and parses unstructured patient records that include medication, electrocardiograms or past medical history. They genotyped 21 single nucleotide polymorphisms (SNPs) in the first 9483 samples accrued into BioVU, and replicated eight of 14 associations with a previously reported odds ratio >1.25, which was significant ($P < 0.05$). These initial sample sizes were small, contributing to the lack of power and differences in identifying these phenotypic groups [43].

EHR-driven phenotypic studies are new and useful source of genetic associations, and can help to identify novel drug targets and to provide validation for existing ones. During the next decade or so, the availability of EHR data enriched with genotypes or genomic sequence information will enable massive *in silico* phenotype–genotype association studies including discovering causal genetic variants for rare Mendelian diseases. Soon we might be able to replicate more GWAS results and find new phenotype–genotype associations. Besides EHR systems implemented at large medical centers, there are also web-based participant-driven studies, such as the Personal Genome Project and 23andMe, in the area of common human traits [44]. Other private health initiatives that are also EHR-related repositories include MicrosoftHealthVault, GoogleHealth [45] and PatientsLikeMe [46]. MicrosoftHealthVault and GoogleHealth let patients organize and share their own health records with family members or health professionals. PatientsLikeMe lets patients with life-changing conditions create and share their health profile, and enables them to find other patients like them enabling them to discuss and share treatment options and concerns. However, the lack of accessible EHR databases enriched with genotypes and genomic sequence information is a limitation for genotype–phenotype association studies. As shown in Table 1, most of the databases do not collect DNA samples from patients because it raises consent and funding issues. In addition, because medical practice varies between countries and cultures, in terms of diagnosis, medications and other more subtle differences, it can be challenging to identify patient cohorts across races and nationalities, and combine their EHR data in a sufficiently homogeneous and unbiased way.

Our hope in the near future, from the data in EHRs and biobanks, is that scientists will be able to understand the relationships between genetic variation in patients and drug response in terms of efficacy and toxicity, which is a crucial step toward personalized medicine [47,48].

Concluding remarks

The burgeoning fields centered on the use of EHRs offer an exciting opportunity for drug discovery, as demonstrated by the above examples of disease relationships, drug–disease interactions and genotype–phenotype associations. Disease relationships identified through EHR mining could shed new light on how diseases are classified and how potential mechanisms and therapies are shared or different. The information on drug–disease interactions can be used to identify new indications for marketed drugs, which will presumably have shorter development times because of their already proven safety. These new indications for drugs might lead to new drug targets and enable the treatment of previously untreatable diseases, or improve existing treatments.

Currently, several challenges hinder even greater research use of EHRs. Clinicians across healthcare systems do not make diagnoses using the same guidelines and classification criteria. EHR data used in the studies were often collected not for research purposes but for billing and administrative purposes. Missing and discrepant data, varying documentation styles and research biases, as well as technical challenges in mining the free-text narrative notes, underscore the difficulties of EHR-based research. Legitimate research results still largely depend on manual review by field staff such as clinicians and nurses. Moreover, current EHR-based research is mainly feasible at large healthcare institutions with adequate funding resources. The magnitude of data also poses serious algorithmic, informatics and information technology challenges. There are legal, ethical, logistical and financial concerns about the accessibility of patient data. Meanwhile, patient populations, standards of care, and standards for data encoding and transfer differ worldwide. Thus, findings from studies mentioned above should be interpreted and generalized with caution.

In conclusion, EHR-based data mining is a rapidly evolving but exciting area. With collaborations among health professionals, epidemiologists, biologists and computational experts in data mining, the impact of EHRs is fast expanding into many aspects of drug discovery. Further development of EHR-based data mining will probably contribute to the quicker identification of safer and more-efficacious drugs and will ultimately have a profound impact on medical care, with untold benefits to patients.

Conflicts of interests

The authors claim no conflicts of interests.

Acknowledgements

We would like to thank Julie Huxley-Jones and Michael Bales for comments and suggestions on the first draft of this manuscript. We are also grateful to Lon Cardon, Vincent Mooser, the reviewers and the Editor for advice.

References

1 Steen, E.B. and Detmer, D.E. (1997) *The Computer Based Patient Record: An Essential Technology for Health Care*. National Academy Press

2 Schoen, C. et al. (2006) *2006 International Health Policy Survey of Primary Care Physicians in Seven Countries*. The Commonwealth Fund

3 Hsiao, C.-J. *et al.* (2009) *Electronic Medical Record/Electronic Health Record Use by Office-based Physicians: United States 2008 and Preliminary 2009*. National Center for Health Statistics

4 Blumenthal, D. and Tavenner, M. (2010) The ‘meaningful use’ regulation for electronic health records. *N. Engl. J. Med.* 363, 501–504

5 Bernstein, K. *et al.* (2005) Modelling and implementing electronic health records in Denmark. *Int. J. Med. Inform.* 74, 213–220

6 Heimly, V. *et al.* (2010) Diffusion and use of Electronic Health Record systems in Norway. *Stud. Health Technol. Inform.* 160, 381–385

7 Shortliffe, E.H. and Cimino, J.J. (2006) *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. Springer

8 Edwards, A. *et al.* (2010) Barriers to cross-institutional health information exchange: a literature review. *J. Healthc. Inf. Manag.* 24, 22–34

9 Vest, J.R. and Jasperson, J. (2010) What should we measure? Conceptualizing usage in health information exchange. *J. Am. Med. Inform. Assoc.* 17, 302–307

10 Balfour, D.C., 3rd *et al.* (2009) Health information technology – results from a roundtable discussion. *J. Manag. Care Pharm.* 15 (Suppl. 1A), 10–17

11 Dean, B.B. *et al.* (2009) Review: use of electronic medical records for health outcomes research: a literature review. *Med. Care Res. Rev.* 66, 611–638

12 Harpe, S.E. (2009) Using secondary data sources for pharmacoepidemiology and outcomes research. *Pharmacotherapy* 29, 138–153

13 Hidalgo, C.A. *et al.* (2009) A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* 5, e1000353

14 Hanauer, D.A. *et al.* (2009) Exploring clinical associations using 'omics' based enrichment analyses. *PLoS One* 4, e5203

15 Jonas, W.B. *et al.* (1996) The effect of niacinamide on osteoarthritis: a pilot study. *Inflamm. Res.* 45, 330–334

16 Ma, A. and Medenica, M. (1983) Response of generalized granuloma annulare to high-dose niacinamide. *Arch. Dermatol.* 119, 836–839

17 Rzhetsky, A. *et al.* (2007) Probing genetic overlap among complex human phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11694–11699

18 Kulkarni, J. *et al.* (2006) A pilot study of hormone modulation as a new treatment for mania in women with bipolar affective disorder. *Psychoneuroendocrinology* 31, 543–547

19 Moretti, M. *et al.* (2011) Tamoxifen effects on respiratory chain complexes and creatine kinase activities in an animal model of mania. *Pharmacol. Biochem. Behav.* 98, 304–310

20 Wang, X. *et al.* (2009) Characterizing environmental and phenotypic associations using information theory and electronic health records. *BMC Bioinformatics* 10 (Suppl. 9), 13

21 Brown, J.S. *et al.* (2007) Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiol. Drug Saf.* 16, 1275–1284

22 Venulet, J. (1988) Possible strategies for early recognition of potential drug safety problems. *Adverse Drug React. Acute Poisoning Rev.* 7, 39–47

23 Wang, X. *et al.* (2009) Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J. Am. Med. Inform. Assoc.* 16, 328–337

24 Bates, D.W. *et al.* (2003) Detecting adverse events using information technology. *J. Am. Med. Inform. Assoc.* 10, 115–128

25 Honigman, B. *et al.* (2001) Using computerized data to identify adverse drug events in outpatients. *J. Am. Med. Inform. Assoc.* 8, 254–266

26 Berlowitz, D.R. *et al.* (2006) Differential associations of beta-blockers with hemorrhagic events for chronic heart failure patients on warfarin. *Pharmacoepidemiol. Drug Saf.* 15, 799–807

27 Wood, L. and Martinez, C. (2004) The general practice research database: role in pharmacovigilance. *Drug Saf.* 27, 871–881

28 Johansson, S. *et al.* (2010) Prospective drug safety monitoring using the UK primary-care General Practice Research Database: theoretical framework, feasibility analysis and extrapolation to future scenarios. *Drug Saf.* 33, 223–232

29 Trifirò, G. *et al.* (2009) Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol. Drug Saf.* 18, 1176–1184

30 Chong, C.R. and Sullivan, D.J., Jr (2007) New uses for old drugs. *Nature* 448, 645–646

31 Chen, H. *et al.* (2005) An epidemiological investigation of off-label anticonvulsant drug use in the Georgia Medicaid population. *Pharmacoepidemiol. Drug Saf.* 14, 629–638

32 Demonaco, H.J. *et al.* (2006) The major role of clinicians in the discovery of off-label drug therapies. *Pharmacotherapy* 26, 323–332

33 Kahan, N.R. *et al.* (2010) Drug use evaluation of tamoxifen focusing on off-label use in a managed care population in Israel. *J. Manag. Care Pharm.* 16, 355–359

34 Marmura, M.J. *et al.* (2010) Electronic medical records as a research tool: evaluating topiramate use at a headache center. *Headache* 50, 769–778

35 Celebisoy, N. *et al.* (2007) Treatment of idiopathic intracranial hypertension: topiramate vs. acetazolamide, an open-label study. *Acta Neurol. Scand.* 116, 322–327

36 Zhou, L. *et al.* (2005) System architecture for temporal information extraction, representation and reasoning in clinical narrative reports. *AMIA Annu. Symp. Proc.* 2005, 869–873

37 Wang, X. *et al.* (2008) Automated knowledge acquisition from clinical narrative reports. *AMIA Annu. Symp. Proc.* 6, 783–787

38 Yao, L. *et al.* (2010) Novel opportunities for computational biology and sociology in drug discovery. *Trends Biotechnol.* 28, 161–170

39 OMOP, (2010) *Applying the OMOP Common Data Model Across Administrative Claims and Electronic Health Records.*

40 Uzuner, O. *et al.* (2008) Identifying patient smoking status from medical discharge records. *J. Am. Med. Inform. Assoc.* 15, 14–24

41 Himes, B.E. *et al.* (2008) Characterization of patients who suffer asthma exacerbations using data extracted from electronic medical records. *AMIA Annu. Symp. Proc.* 6, 308–312

42 Clayton, E.W. *et al.* (2010) Confronting real time ethical, legal, and social issues in the Electronic Medical Records and Genomics (eMERGE) Consortium. *Genet. Med.* 12, 616–620

43 Ritchie, M.D. *et al.* (2010) Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* 86, 560–572

44 Eriksson, N. *et al.* (2010) Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* 6, e1000993

45 Do, N.V. *et al.* (2011) The military health system's personal health record pilot with Microsoft HealthVault and Google Health. *J. Am. Med. Inform. Assoc.* 18, 118–124

46 Frost, J. and Massagli, M. (2009) PatientsLikeMe the case for a data-centered patient community and how ALS patients use the community to inform treatment decisions and manage pulmonary health. *Chron. Respir. Dis.* 6, 225–229

47 McCarty, C.A. and Wilke, R.A. (2010) Biobanking and pharmacogenomics. *Pharmacogenomics* 11, 637–641

48 Wilke, R.A. *et al.* (2011) The emerging role of electronic medical records in pharmacogenomics. *Clin. Pharmacol. Ther.* 89, 379–386

49 Bruggenjurgen, B. *et al.* (2007) Utilisation of medical resources of patients with pain undergoing an outpatient opioid therapy. *Gesundheitswesen* 69, 353–358

50 Waterstone, J. and Parsons, J. (1992) Endometrial stromal sarcoma two years after a successful *in vitro* fertilization treatment cycle. *Hum. Reprod.* 7, 72

51 Mullins, I.M. *et al.* (2006) Data mining and clinical data repositories: insights from a 667,000 patient data set. *Comput. Biol. Med.* 36, 1351–1377