Published in final edited form as:

J Surg Res. 2020 February ; 246: 599–604. doi:10.1016/j.jss.2019.09.053.

What Can We Learn About Drug Safety and Other Effects in the Era of Electronic Health Records and Big Data That We Would Not Be Able to Learn From Classic Epidemiology?

Ali Zarrinpar, MD, PhD^{a,*}, Ting-Yuan David Cheng, PhD^b, Zhiguang Huo, PhD^c

^aDepartment of Surgery, College of Medicine, University of Florida, Gainesville, Florida

^bDepartment of Epidemiology, College of Public Health & Health Professions and College of Medicine, University of Florida, Gainesville, Florida

^cDepartment of Biostatistics, College of Public Health & Health Professions and College of Medicine, University of Florida, Gainesville, Florida

Abstract

As more and more health systems have converted to the use of electronic health records, the amount of searchable and analyzable data is exploding. This includes not just provider or laboratory created data but also data collected by instruments, personal devices, and patients themselves, among others. This has led to more attention being paid to the analysis of these data to answer previously unaddressed questions. This is especially important given the number of therapies previously found to be beneficial in clinical trials that are currently being re-scrutinized. Because there are orders of magnitude more information contained in these data sets, a fundamentally different approach needs to be taken to their processing and analysis and the generation of knowledge. Health care and medicine are drivers of this phenomenon and will ultimately be the main beneficiaries. Concurrently, many different types of questions can now be asked using these data sets. Research groups have become increasingly active in mining large data sets, including nationwide health care databases, to learn about associations of medication use and various unrelated diseases such as cancer. Given the recent increase in research activity in this area, its promise to radically change clinical research, and the relative lack of widespread knowledge about its potential and advances, we surveyed the available literature to understand the strengths and limitations of these new tools. We also outline new databases and techniques that are available to researchers worldwide, with special focus on work pertaining to the broad and rapid monitoring of drug safety and secondary effects.

Keywords

Electronic health record; Big data; Drug safety; Health care database; Cancer risk

* *Corresponding author.* Department of Surgery, University of Florida, College of Medicine, 1600 SW Archer Road, Gainesville, FL 32610-0118. Tel.: +1 3102679610; fax: +1 3102673590., ali.zarrinpar@surgery.ufl.edu (A. Zarrinpar).

Disclosure

The authors have no financial or personal relationships with other people or organizations that could potentially and inappropriately influence this work and its conclusions.

Introduction

As human approaches to science and knowledge have evolved and expanded from the empirical (describing natural phenomena) to theoretical (generating and using models to generalize observations) to computational (simulating complex phenomena), we have taken increasing advantage of more powerful methods of capturing and processing expanding amounts of information. This process, however, appears to be expanding exponentially, and to learn from all this new information, scientists have begun to consolidate the theoretical, experimental, and computational data. Data are collected by instruments or created by simulations, they are processed by software, the information is stored, and these databases are analyzed by sophisticated data management and statistics. All this represents a significantly, if not fundamentally, new approach toward the generation of knowledge, with health care being both a driver and a beneficiary of new methodologies.

Increases in computational power, data storage capacity, public interest, and the political will to support the wider establishment and utilization of the electronic health records (EHRs) worldwide has brought with it a veritable flood of individual health information. In parallel, advances in big data have dramatically changed not just what is considered data but also how data are analyzed and for what purposes.¹ Beyond the business world, more recent forays have been made into clinical research and even clinical practice.

While much of the promise of the EHR revolves around the ability to rapidly acquire, record, provide, process, synthesize, and transmit data to facilitate day-to-day patient care, clinician scientists who have previously relied upon classic clinical epidemiology can now have access to this tremendous resource. The increasing digitization of the health record and the expanding incorporation of large amounts of disparate data from growing pools of people also allow for discoveries that are too stochastic or rare to be found by individual observations or even by moderately sized clinical studies. Many groups have begun to mine large data sets, including nationwide health care databases, to learn about associations of medication use and various unrelated diseases such as cancer, as well as unforeseen safety concerns. These groups have thus far correlated the use of a variety of drugs with increased incidence of cancers, aspirin and statins with decreased cancer risk, and benzodiazepines with increased dementia.^{2–7} Data such as these have also made their way into artificial intelligence models that not only aid in predicting surgical outcomes but also can make recommendations with regard to surgical decision-making.⁸ Herein, we intend to survey the available literature, understand the strengths and limitations of these new tools, and outline new databases and techniques that are available to researchers worldwide, especially in work pertaining to the broad and rapid monitoring of drug safety and secondary effects.

History of pharmacoepidemiologic studies and methods

Traditional biostatistics and pharmacoepidemiologic studies have relied heavily on simple distribution statistics and multivariable regression. This is due to the fact that the generation and analysis of such data are difficult and expensive. Efficiency in both the acquisition of the data and their analyses were necessary. This state has evolved over the years, however, to one in which practically every device in the hospital and clinic generates data, from

ventilators to fluid pumps. Furthermore, analytic methods have also evolved because computation is cheap and readily available. This allows the ability to “let patterns emerge from the data” and therefore to test multiple sometimes interrelated hypotheses together or in rapid succession.⁹

The implications of these new techniques are that we can now do analyses with increases in the “three V’s”: variety, volume, and velocity.^{10,11} Increased variety, meaning data from different sources and more variables per subject or observation, greatly complicates variable selection and selectivity but allow for the assessment of complex interaction. Increased volume, simply meaning more observations, allows for increased power to detect small or unexpected associations, with the understanding that this may mean losing meaningful relevance or improperly inferring causation. Increased velocity in collecting and analyzing data can lead to dynamic interventions, rapid adaptation, and discovery and can immediately impact the public health and safety. Just as the cost of sequencing has decreased by orders of magnitude over the last decade and the number of transistors in an integrated circuit has doubled every 18 mo, one can imagine that while health care is not currently equipped to respond appropriately to rapid data analysis at this time, innovations will continue to augment our abilities to do so.

Availability of new databases

Since John Graunt used burial data to perform one of the world’s first recorded epidemiologic studies in 1662, health records have been repurposed from their initial purposes (such as care delivery and billing) into use for the study of medicine and medical care. In addition to EHR, population-based registries and administrative claims databases have been widely utilized (Table). Many countries with integrated health care delivery systems in Europe and Asia now have large interlinked data sets containing large number of patients with longitudinal data of a variety of types that are connected to biobanks, outcome measures, and even social media.^{12,13} These previously unavailable databases can even be combined to provide population diversity and better signal for a more rapid detection of associations.¹⁴ For example, the use of a larger database could have accelerated the detection of the adverse effects of rofecoxib by an order of magnitude (from 5 y to 3 mo).¹⁵ Exceedingly rare outcomes of interest could also be studied in this manner, such as the safety of the use of selective serotonin reuptake inhibitors by pregnant women.¹⁶

Some national databases have been designed with bioinformatics in mind from the initial stages. In Taiwan, for example, 99.9% of the 23 million people are covered by and incorporated into the National Health Insurance (NHI) system and thus part of the electronic medical record.¹⁷ The NHI program, established more than 20 y ago, delivers universal coverage through a “government-run, single-payer compulsory insurance plan to centralize the disbursement of health care funding.”¹⁸ Almost immediately after the establishment of the NHI, several data centers were established to mine its data. Thus far the NHI has produced greater than 1000 papers.

Indeed, while data collected and maintained by public institutions in single-payer settings have the advantage of lifelong follow-up and universal health care access, databases such as

US claims databases can illuminate the effects of demographic and economic disparities and also provide opportunities for cross-validation.^{14,19} For the study of therapeutics, though, such databases run into problems such as concerns about data validity, misclassifications of variables, lack of detailed clinical information or human-mediated verification of data validity, and thus the loss of the ability to account for confounding factors.²⁰

Availability of new methodology

An example of new methodology being applied to monitoring drug safety, the FDA, spurred by Congress, launched an initiative to use health care data including claims databases, EHRs, and registries, to evaluate safety issues of medications and related products. The current standard relies on voluntary submissions of Adverse Event Reports, which then have to be analyzed statistically for safety signals.²¹ Liu *et al.* have built a methodology to use texts of clinical notes and incorporate temporal patterns of medication administration and adverse effects to discover new adverse events.

Classically, the gold standard of drug studies has been the randomized controlled trials (RCTs). Clinical trials with adaptive designs enjoy increased popularity due to the capability to adjust the recruitment allocation by utilizing the information already collected in the trial. Common adaptive designs include re-estimating the sample size to ensure power, shifting the allocation ratio to more promising doses or treatments, restricting the recruitment toward the most beneficial subpopulation, and terminating the trial for success or lack of efficacy.²² Compared with the standard RCTs, adaptive designs are more efficient and ethical because they usually consume less time, money, and participants,²³ though challenges in randomization schedule and statistical inference still exist.²⁴ Such practice has been included in the FDA draft guidance for industry entitled “Adaptive Designs for Clinical Trials of Drugs and Biologics” (FDA-2018-D-3124).

More recently, with the ready availability of longitudinal health care databases, especially as they are most often the only available data source for the study of medications already in clinical use, regulatory agencies have relied upon their analysis for postapproval monitoring. To combat the various confounding factors in this type of analysis, new methodologies have been developed including highdimensional propensity score approaches.²⁵ This technique involves the identification and subsequent prioritization of covariates, which leads to potential estimation of causal treatment effect under certain statistical assumptions. This allows for the use of large numbers of covariates in a variety of settings.

Another exciting new possibility that takes advantage of large populations and big data is molecular pathological epidemiology.²⁶ This field incorporates individual differences, including genetic and environmental heterogeneity, into epidemiologic analysis using molecular pathology. Instead of assuming that any given disease manifests in the same way in different patients, molecular pathological epidemiology is founded on the idea that every disease is unique. For example, ever since Perou *et al.*,²⁷ among others identified clinically meaningful subtypes of breast cancer (i.e., Luminal A, Luminal B, Her2-enriched, Basal-like and Normal-like) using gene expression microarray data, these subtypes have been increasingly utilized to stratify patients to receive their optimized treatment.²⁸ But molecular

pathological epidemiology is especially important in pharmacoepidemiology, that is, the study of the effects of drugs on diseases and their potential side effects in humans.^{29,30} Similar efforts have been made in the discovery of other powerful tools to reduce confounding and other biases inherent in epidemiologic predictions³¹ and in complex-trait prediction.³²

Various groups have applied such techniques specifically to associations between medications and their side effects. These efforts have been instigated by a growing consensus that many adverse effects of drugs are being detected too late, after many people have already suffered. Improvements in drug safety monitoring would include not just the ability to investigate a wide variety of possible adverse drug effects but also to find them at least close to, if not in, real time. Trifiro *et al.* made some progress in this by developing a method to systematically identify events important to pharmacovigilance.^{15,33} Having such a list of events would have the additional effect of decreasing spurious findings. Wright *et al.*³⁴ describe an approach to the use of a data mining technique, termed association rule mining, and found it to be useful in detecting clinically accurate associations between medications, laboratory results, and problems.

Selected recent studies in drug toxicity and pharmacovigilance

An example of a class of medication that are both widely used in the general population and highly studied (likely as a result of their widespread use) is the statin family. Statins comprise a group of the most commonly prescribed drugs, used to treat hyperlipidemia. They are known to have many side effects, including myopathy. Since their introduction, however, they have been hypothesized to have other previously unknown toxicities and also many overall benefits. One example of a toxicity was its presumed teratogenicity. Multiple small studies including those from registries, small cohort studies, and case reports failed to show a consistent association or lack thereof. Bateman *et al.*³⁵ used the Medicaid Analytic eXtract to show that not only statins did not themselves cause teratogenic effects but also that the previously found associations were likely due to the young mothers having other underlying disease, such as pre-existing diabetes, that can contribute to the congenital malformations.

Large health care claims databases can be used to parse out drug effects and/or toxicities in specific patient populations due to the large numbers of people in the database. El-Refai *et al.*³⁶ were able to show that statins protect against venous thromboembolism, not just in the general population but also in cancer patients.

Another widely used drug, acetaminophen, has been studied in several epidemiologic and clinical studies to determine its relationship with analgesic nephropathy without much success. Because of the difficulty in objective and unbiased measure of acetaminophen use due to issues such as recall bias and exposure misclassification, Kelkar *et al.* utilized an administrative claims-based approach. What they found was that acute use of high doses of acetaminophen was in fact associated with renal disease, but chronic use was not associated with an increased risk.³⁷

The question of power of the analysis to delineate causes of rare events in a rare disease is an important one and difficult to answer using standard studies that rely on spontaneous reports to the FDA's Adverse Event Reporting System. For example, reports of lymphoma in patients with juvenile idiopathic arthritis (JIA) treated with tumor necrosis factor (TNF) inhibitors led to the placement of a black box warning implicating TNF inhibitors with increasing the risk of malignancy. This problem is complicated by a variety of factors including i) reliance on voluntary reporting and the resulting underestimation of true number of events and overestimation of well-publicized events, ii) inaccurate drug dose or exposure estimates, and iii) limited knowledge about the background rate of rare events in a rare population. Beukelman *et al.* addressed this question using US national data containing administrative claims records from Medicaid programs. This allowed them to find cohorts of children with JIA and two comparator cohorts of children with other nonimmune/inflammatory chronic diseases (attention-deficit/hyperactivity disorder and asthma). They were also able to find whether these patients were treated with TNF inhibitors or not. They concluded that JIA itself appeared to increase rates of malignancy and that TNF inhibitors had no significant association.³⁸

While this study was unable to evaluate the doses of TNF inhibitors and whether they mattered, large database studies can parse out the effect of dosage on drug toxicities. For example, Tsai *et al.*³⁹ found a robust dose effect comparing low- and high-dose metoclopramide on the development of Parkinsonism, thus supporting a proposed restriction on the duration of length of treatment.

Jack Li's group at the Taipei Medical University has taken advantage of the Taiwan National Health Insurance Database and the power of large amounts of longitudinal, methodically collected data to study whether the use of various medications or devices protects against or causes cancer.³ These associations range from benzodiazepines^{2,4} to statins⁷ to levothyroxine⁴⁰ to cell phone use.⁴¹ They are not alone in that other groups and countries with national health databases have similarly examined other drugs and their potential toxicities and/or associated complications.^{42–44}

What is novel and could not be done otherwise?

There is a fresh perspective and energy in the field that is driving innovation. The novelty comes in its opposition to RCTs. In RCTs the question is “can this treatment work?” or “is it efficacious?” The patient populations are relatively small, highly selected, and not likely to be applicable to a larger population; the comparators are either placebo or a predetermined “routine care”; the outcome measures are easy to measure surrogates but not necessarily representative of what patients or physicians want; follow-up time is relatively short; speed of the study is slow; and per patient cost is high.⁴⁵

In the era of EHRs and big data, studies are essentially comparative effectiveness research. That is, they compare the benefits and harms of a variety of methods to prevent, diagnose, treat, and monitor clinical conditions. The question is “how well does this treatment work in clinical practice?” The patient populations are very large and more representative of actual

practice; the comparators are active and variable; outcome measures are patient centered or global; follow-up time is long; cost is low, especially per patient; and the speed is fast.

Ethical concerns and potential solutions

Among the rapid development of big data research using EHR, the number one ethical concern is patient privacy. By far, human research is tightly regulated under the Health Insurance Portability and Accountability Act (HIPAA) to protect patient information. However, patient advocates and research community have called for relaxing regulation to facilitate data sharing.^{46,47} In many institutions such as NCI-designated cancer centers, patient consent has been built into the patient's check-in process to ensure free sharing of medical information and biospecimens, as they can greatly facilitate and expedite patient care. With appropriate measures of protection in data and information, a waiver of patients' consents can be granted by an institutional review board (IRB) at local institutions. A successful model is OneFlorida Clinical Research Consortium that already gathers health information of a third of Floridians from different health systems.⁴⁸ OneFlorida provides deidentified data sets for research once a research protocol is approved by OneFlorida committees and individual institution's IRB. Also, patient counts based on almost all variables in the EHR are available without IRB approval, which greatly facilitates project planning. Aside from using existing data, patient re-contact has been a frontier facilitated by hospitals asking patients' consent upon admission.⁴⁹ With regard to the use of biospecimens for research, a well-defined, feasible protocol to deidentify blood or tissue from multiple institutes is key to promoting big data approach in biospecimens without patient consent.

Conclusion

While RCTs will not be supplanted any time soon and maintain their deserved place in safety and efficacy studies, we will be seeing more efficacy and safety studies coming from EHR/big data studies. To optimize this process, it becomes essential to ensure data validity and avoid misclassification. One way to improve this would be the use of well-designed data management software, including REDCap (Research Electronic Data Capture) databases designed specifically for clinical and translational research. Another would be using EHR databases from platforms such as Epic, where the validity of the data is important not just for research or billing purposes but also for actual clinical care, and thus would have more human-mediated editing and validation. These would be a more monitored form of a large database. In the next several years, it is likely that widely held concepts in medicine and medication efficacy and safety will be tested using these new techniques. While it is easy to imagine that many dogmas will stand the test of time, the ones that will fail are quite unpredictable.

Acknowledgment

Authors' contributions: A.Z. (NIH/NIDDK K08DK113244), T-Y.D.C., and Z.H. contributed to the collection of data and background, writing, and revision of the manuscript.

REFERENCES

1. Shortreed SM, Cook AJ, Coley RY, Bobb JF, Nelson JC. Challenges and opportunities for using big health care data to advance medical science and public health. *Am J Epidemiol.* 2019;188:851–861. [PubMed: 30877288]
2. Iqbal U, Nguyen PA, Syed-Abdul S, et al. Is long-term use of benzodiazepine a risk for cancer? *Medicine (Baltimore).* 2015;94:e483. [PubMed: 25674736]
3. Iqbal U, Hsu CK, Nguyen PA, et al. Cancer-disease associations: a visualization and animation through medical big data. *Comput Methods Programs Biomed.* 2016;127:44–51. [PubMed: 27000288]
4. Iqbal U, Jian WS, Huang CW, Inayat A, Li YC. Do all hypnotic and sedatives have risk for cancer? *Sleep Med.* 2016;20:170. [PubMed: 26323647]
5. Iqbal U, Yang HC, Jian WS, Yen Y, Li YJ. Does aspirin use reduce the risk for cancer? *J Investig Med.* 2017;65:391–392.
6. Islam MM, Iqbal U, Walther B, et al. Benzodiazepine use and risk of dementia in the elderly population: a systematic review and meta-analysis. *Neuroepidemiology.* 2016;47:181–191. [PubMed: 28013304]
7. Islam MM, Yang HC, Nguyen PA, et al. Exploring association between statin use and breast cancer risk: an updated meta-analysis. *Arch Gynecol Obstet.* 2017;296:1043–1053. [PubMed: 28940025]
8. Loftus TJ, Upchurch GR, Bihorac A. Use of artificial intelligence to represent emergent systems and augment surgical decision-making. *JAMA Surg.* 2019.
9. Iwashyna TJ, Liu V. What's so different about big data?. A primer for clinicians trained to think epidemiologically. *Ann Am Thorac Soc.* 2014;11:1130–1135. [PubMed: 25102315]
10. Mooney SJ, Westreich DJ, El-Sayed AM. Commentary: epidemiology in the era of big data. *Epidemiology.* 2015;26:390–394. [PubMed: 25756221]
11. Baro E, Degoul S, Beuscart R, Chazard E. Toward a literature-driven definition of big data in healthcare. *Biomed Res Int.* 2015;2015:639021. [PubMed: 26137488]
12. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562:203–209. [PubMed: 30305743]
13. Convertino I, Ferraro S, Blandizzi C, Tuccori M. The usefulness of listening social media for pharmacovigilance purposes: a systematic review. *Expert Opin Drug Saf.* 2018;17:1081–1093. [PubMed: 30285501]
14. Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. *Clin Epidemiol.* 2017;9:245–250. [PubMed: 28490904]
15. Trifiro G, Coloma PM, Rijnbeek PR, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *J Intern Med.* 2014;275:551–561. [PubMed: 24635221]
16. Furu K, Kieler H, Haglund B, et al. Selective serotonin reuptake inhibitors and venlafaxine in early pregnancy and risk of birth defects: population based cohort study and sibling design. *BMJ.* 2015;350:h1798. [PubMed: 25888213]
17. Li YC, Yen JC, Chiu WT, et al. Building a national electronic medical record exchange system - experiences in Taiwan. *Comput Methods Programs Biomed.* 2015;121:14–20. [PubMed: 26001420]
18. Hsing AW, Ioannidis JP. Nationwide population science: lessons from the Taiwan national health insurance research database. *JAMA Intern Med.* 2015;175:1527–1529. [PubMed: 26192815]
19. Ehrenstein V, Sørensen HT, Bakkevig LS, Pedersen L. Medical databases in studies of drug teratogenicity: methodological issues. *Clin Epidemiol.* 2010;2:37–43. [PubMed: 20865101]
20. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol.* 2005;58:323–337. [PubMed: 15862718]
21. Liu Y, Lependu P, Iyer S, Shah NH. Using temporal patterns in medical records to discern adverse drug events from indications. *AMIA Jt Summits Transl Sci Proc.* 2012;2012:47–56. [PubMed: 22779050]

22. Chow SC, Chang M. Adaptive design methods in clinical trials - a review. *Orphanet J Rare Dis*. 2008;3:11. [PubMed: 18454853]
23. Pallmann P, Bedding AW, Choodari-Oskooei B, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med*. 2018;16:29. [PubMed: 29490655]
24. Chow SC, Corey R. Benefits, challenges and obstacles of adaptive clinical trial designs. *Orphanet J Rare Dis*. 2011;6:79. [PubMed: 22129361]
25. Schneeweiss S Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clin Epidemiol*. 2018;10:771–788. [PubMed: 30013400]
26. Hamada T, Keum N, Nishihara R, Ogino S. Molecular pathological epidemiology: new developing frontiers of big data science to study etiologies and pathogenesis. *J Gastroenterol*. 2017;52:265–275. [PubMed: 27738762]
27. Perou CM, Sørli T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406:747–752. [PubMed: 10963602]
28. Duffy MJ, Harbeck N, Nap M, et al. Clinical use of biomarkers in breast cancer: updated guidelines from the european group on Tumor Markers (EGTM). *Eur J Cancer*. 2017;75:284–298. [PubMed: 28259011]
29. Fink SP, Yamauchi M, Nishihara R, et al. Aspirin and the risk of colorectal cancer in relation to the expression of 15-hydroxyprostaglandin dehydrogenase (HPGD). *Sci Transl Med*. 2014;6:233re232.
30. Chan AT, Ogino S, Fuchs CS. Aspirin and the risk of colorectal cancer in relation to the expression of COX-2. *N Engl J Med*. 2007;356:2131–2142. [PubMed: 17522398]
31. Toh S Pharmacoevidence in the era of real-world evidence. *Curr Epidemiol Rep*. 2017;4:262–265. [PubMed: 29204331]
32. de Los Campos G, Vazquez AI, Hsu S, Lello L. Complex-trait prediction in the Era of big data. *Trends Genet*. 2018;34:746–754. [PubMed: 30139641]
33. Trifiro G, Pariente A, Coloma PM, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoevidence Drug Saf*. 2009;18:1176–1184. [PubMed: 19757412]
34. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. *J Biomed Inform*. 2010;43:891–901. [PubMed: 20884377]
35. Bateman BT, Hernandez-Diaz S, Fischer MA, et al. Statins and congenital malformations: cohort study. *BMJ*. 2015;350:h1035. [PubMed: 25784688]
36. El-Refai SM, Black EP, Adams VR, Talbert JC, Brown JD. Statin use and venous thromboembolism in cancer: a large, active comparator, propensity score matched cohort study. *Thromb Res*. 2017;158:49–58. [PubMed: 28822240]
37. Kelkar M, Cleves MA, Foster HR, et al. Acute and chronic acetaminophen use and renal disease: a case-control study using pharmacy and medical claims. *J Manag Care Pharm*. 2012;18:234–246. [PubMed: 22468732]
38. Beukelman T, Haynes K, Curtis JR, et al. Rates of malignancy associated with juvenile idiopathic arthritis and its treatment. *Arthritis Rheum*. 2012;64:1263–1271. [PubMed: 22328538]
39. Tsai SC, Sheu SY, Chien LN, et al. High exposure compared with standard exposure to metoclopramide associated with a higher risk of parkinsonism: a nationwide population-based cohort study. *Br J Clin Pharmacol*. 2018;84:2000–2009. [PubMed: 29745438]
40. Wu CC, Yu YY, Yang HC, et al. Levothyroxine use and the risk of breast cancer: a nation-wide population-based case-control study. *Arch Gynecol Obstet*. 2018;298:389–396. [PubMed: 29961136]
41. Hsu MH, Syed-Abdul S, Scholl J, et al. The incidence rate and mortality of malignant brain tumors after 10 years of intensive cell phone use in Taiwan. *Eur J Cancer Prev*. 2013;22:596–598. [PubMed: 23591455]
42. Frisk G, Ekberg S, Lidbrink E, et al. No association between low-dose aspirin use and breast cancer outcomes overall: a Swedish population-based study. *Breast Cancer Res*. 2018;20:142. [PubMed: 30458873]

43. Bruun SB, Petersen I, Kristensen NR, Cronin-Fenton D, Pedersen AB. Selective serotonin reuptake inhibitor use and mortality, postoperative complications, and quality of care in hip fracture patients: a Danish nationwide cohort study. *Clin Epidemiol*. 2018;10:1053–1071. [PubMed: 30214311]
44. Christensen J, Grønberg TK, Sørensen MJ, et al. Prenatal valproate exposure and risk of autism spectrum disorders and childhood autism. *JAMA*. 2013;309:1696–1703. [PubMed: 23613074]
45. Mansournia MA, Higgins JP, Sterne JA, Hernán MA. Biases in randomized trials: a conversation between Trialists and epidemiologists. *Epidemiology*. 2017;28:54–59. [PubMed: 27748683]
46. Spector-Bagdady K, Fernandez Lynch H, Brenner JC, Shuman AG. Biospecimens, Research consent, and distinguishing cell line research. *JAMA Oncol*. 2019;5:406–410. [PubMed: 30703213]
47. Weng C, Appelbaum P, Hripcsak G, et al. Using EHRs to integrate research with patient care: promises and challenges. *J Am Med Inform Assoc*. 2012;19:684–687. [PubMed: 22542813]
48. Shenkman E, Hurt M, Hogan W, et al. OneFlorida clinical research Consortium: linking a clinical and translational science institute with a community-based distributive medical education model. *Acad Med*. 2018;93:451–455. [PubMed: 29045273]
49. Flood-Grady E, Clark VC, Bauer A, et al. Evaluating the efficacy of a registry linked to a consent to re-contact program and communication Strategies for recruiting and enrolling participants into clinical trials. *Contemp Clin Trials Commun*. 2017;8:62–66. [PubMed: 29503877]

Table -

Emerging databases and approaches and their strengths and challenges.

Database and approach	Strengths	Challenges
Administrative claims databases	Highly structured; long history of providing comparative effectiveness and drug safety data	Limited by payer's system; acquiring unmeasured assessment and behavioral variables; inflated type I error rate if no strong a priori hypothesis
Population-based registries and biobanks	Nationwide representativeness; availability of genomic and phenotypic data	Limited behavioral variables
Electronic health records	Detailed treatment, imaging, and prescribing data; availability of provider notes; common data models (CDM) improve data quality	Selective measurement; selection bias; the need for integration of multiple providers and systems to avoid loss of follow-up
Adaptive designs for clinical trials	More efficient and ethical than standard randomized controlled trials	Improving interpretation, communication, and implementation of the results
Automated data adaptive analytics for EHR	Improving confounding adjustment and causal inference	Proper use of machine learning and its transparency
Molecular pathological epidemiology	Elucidating biological mechanisms and new treatment strategies	Tissue availability; laboratory assay development and cost