Drug Saf (2015) 38:601–610
DOI 10.1007/s40264-015-0305-9

CURRENT OPINION

Electronic Health Data for Postmarket Surveillance: A Vision Not Realized

Thomas J. Moore[1,2] · **Curt D. Furberg**[3]

Published online: 30 May 2015

Abstract What has been learned about electronic health data as a primary data source for regulatory decisions regarding the harms of drugs? Observational studies with electronic health data for postmarket risk assessment can now be conducted in Europe and the US in patient populations numbering in the tens of millions compared with a few hundred patients in a typical clinical trial. With standard protocols, results can be obtained in a few months; however, extensive research published by scores of investigators has illuminated the many obstacles that prevent obtaining robust, reproducible results that are reliable enough to be a primary source for drug safety decisions involving the health and safety of millions of patients. The most widely used terminology for coding patient interactions with medical providers for payment has proved ill-suited to identifying the adverse effects of drugs. Directly conflicting results were reported in otherwise similar patient health databases, even using identical event definitions and research methods. Evaluation of some accepted statistical methods revealed systematic bias, while others appeared to be unreliable. When electronic health data studies detected no drug risk, there were no robust and accepted standards to judge whether the drug was unlikely to cause the adverse effect or whether the study was incapable of detecting it. Substantial investment and careful thinking is needed to improve the reliability of risk assessments based on electronic health data, and current limitations need to be fully understood.

Key Points

Electronic health data for postmarket surveillance became a key element in the new paradigm for drug regulation, which involved fewer and smaller clinical trials prior to marketing approval.

The research programs and pilot systems created to study harms of licensed drugs proved largely unable to provide credible evidence of new, unsuspected drug adverse effects, and conflicting and contradictory results when seeking to confirm known harms.

Major problems included a limited underlying terminology, few validation studies, and the need for additional statistical standards for these complex data.

✉ Thomas J. Moore
tmoore@ismp.org

[1] Institute for Safe Medication Practices, 101 N. Columbus St, Suite 410, Alexandria, VA 22214, USA

[2] Department of Epidemiology and Biostatistics, George Washington University Milken Institute School of Public Health, Washington, DC, USA

[3] Division of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, NC, USA

1 Introduction

Over nearly a decade, regulators, drug developers, and the epidemiology community developed an ambitious new vision of postmarket surveillance. The globalization of health information in digital form, powerful new statistical tools, and ever-expanding computing capabilities offered

the promise of learning more about the harms of pharmaceutical drugs than traditional methods, as well as doing it faster, and at lower cost. Extensive research programs were conceived to determine the best methodological approaches; health databases with tens of millions of patients were created, and both known and suspected adverse drug effects were studied in depth. Our objective was to assess what has been learned about the value of electronic health data as a primary data source for regulatory decisions regarding the harm of drugs.

1.1 The New Paradigm Emerges

"The days of largely relying on clinical trials pre-authorization and spontaneous reports of suspected adverse reactions post-authorization are over" declared two senior officials of the European Medicines Agency (EMA) in a recent summary of the new paradigm for drug regulation [1]. The new paradigm policies described by both European and US regulators call for shorter, smaller, more flexible clinical trials prior to approval but additional clinical studies, risk management plans, and more intensive postmarket surveillance post approval [2]. One key element was electronic health data to permit better and faster postmarket surveillance. In terms of a timeline, the first blueprints for the new paradigm began to emerge around 2006 with the launch of a study from the US Institute of Medicine [3], with implementation of various elements continuing at present [4].

In addition to specific regulatory needs, 'big data' combined with sophisticated statistical methods has also become a glamorous tool for investigating a complex, interconnected world; celebrated applications include predicting election outcomes, targeting financial markets for split-second trading, and increasing retail sales through tracking minute-to-minute movements of millions of customers through large stores. Furthermore, big data approaches offered results that were fast, low cost because they utilized existing data, and appeared capable of uncovering statistical relationships that no one had previously suspected. It was inevitable that health outcomes and drug risks could and should be included under the umbrella of big data.

1.2 The Need for High Standards

It was one thing to describe 'active surveillance' and 'big data' and advances in 'regulatory science' in glowing terms, and quite another to extract from billions of disparate health transactions compelling risk assessments of sufficient validity that they could support regulatory decisions about the adverse effects of drugs. With hundreds of millions of dollars in drug revenue and the health of millions of patients at stake, observational studies regarding drug risks required robust, reproducible assessments.

To the extent the new paradigm envisioned using electronic health data studies as a backstop for fewer and shorter randomized clinical trials, the quality and validity standards needed to be higher still. It was one thing to produce some interesting new perspectives on drug risks gleaned from health insurance claims, or to refine risk assessments of findings originating from other data sources, and quite another to imagine these data could provide safety assurances comparable to the results of well-conducted randomized clinical trials.

2 Implementing the New Paradigm

Three seminal research projects were conceived to build a bridge between the vision and a usable reality using electronic health data for postmarket surveillance [5]. In the US, the FDA, the National Institutes of Health, and the drug industry lobbying organization [Pharmaceutical Research and Manufacturers of America (PhRMA)] created OMOP (the Observational Medical Outcomes Partnership) to study and validate study methods [6]. In Europe, the EMA joined 33 partners (mainly drug companies and research institutions) in a 5-year project called PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium). Without waiting for these research-oriented efforts to complete their multi-year research programs, in 2009 the FDA launched a national medical product safety surveillance system called the Sentinel Initiative [7]. At present, all three projects have completed their initial cycles.

In addition to these research initiatives, the basic idea of observational studies in electronic health information was not new. For many years, observational studies had been conducted in the national health records of Denmark and the UK, in the Canadian province of Saskatchewan health data, and in US Medicaid programs that provide state-level medical insurance for the poor and disabled.

3 The Observational Medical Outcomes Partnership (OMOP) Experiment

The most extensive and systematic investigation of the future of drug risk assessment using electronic health data came through the 5-year OMOP initiative [8]. Researchers assembled five different electronic databases—four with insurance claims data and one with electronic health records. The populations were diverse, including older patients in the US Medicare program (4.6 million persons), a low-income population in state-level Medicaid programs

(10.8 million), and families covered by employer-based insurance (46.5 million) [9].

The investigators selected four health outcome events to study [9]. All were acute events likely to require hospitalization or other intensive medical treatment, and therefore relatively easy to identify in databases of health insurance claims. The outcome events—acute liver failure, acute kidney failure, acute myocardial infarction, and upper gastrointestinal bleed—have been previously linked to various drugs as causative agents.

To test seven different analytical methods, the project created 165 positive controls, pairs of drug–health outcomes that were clearly associated in the literature, and then created 234 negative controls identifying drug–event pairs where the available evidence showed no association. The different statistical methods (such as case control or Bayesian data-mining algorithms) could then be evaluated based on their ability to confirm the known associations, but find no elevated risk where no association was suspected.

The obstacles that this large-scale analytical project had to overcome were numerous. Even with only four adverse event outcomes under investigation, an extensive literature search and evaluation was required to identify the 165 drug–event pairs in which the drug was clearly implicated [10]. The search had to be repeated to document the 234 cases where no association was suspected; proving a negative is always an uncertain task because undocumented adverse effects could well exist. Outcome events identified through diagnosis codes in disparate databases had to be investigated and prevalence varied substantially by event definition. Each of seven statistical methods had to be standardized, and each had a series of two to six different parameters.

3.1 OMOP Results

Viewed independently, the main message of OMOP was that study results were variously unreliable, inconsistent, and sometimes contradictory, affecting every one of the multiple major parameters in the study. Methods differed substantially in their ability to accurately identify the 165 'true' drug–event relationships [9]. In these studies, two methods—Bayesian and frequentist disproportionality methods for safety signaling—failed to discriminate between true positives and assumed negative controls [11]. Self-controlled methods performed better than case controls and new user cohorts, even though the latter two methods are widely used in other observational studies [12]. The validity of a single method depended heavily on the specific study parameters, which were numerous for each method. An optimal set of parameters for one adverse event turned out to be suboptimal for other adverse events. Also, a higher ranked method, optimized, produced different and sometimes contradictory results among the five different databases for the same adverse event [13]. Finally, some methods revealed substantial evidence of systematic bias, finding statistically significant associations between drugs and events among the negative controls where no relationship was thought to exist.

These issues can be illustrated in the results for the self-controlled cohort method [12]. In this method, the event rates in patients during a time period when they are not exposed to the target drug are compared with the rates 30 or 180 days after exposure. The approach was among the most successful overall in discriminating between true drug–event relationships and the negative controls; however, more specific findings were not reassuring. For the drug–acute liver failure event pair, the approach successfully identified isoniazid, which has a boxed warning for fatal hepatitis, but not for erythromycin, which has a prominent hepatotoxicity warning. For the negative controls where no relationship was suspected, it detected no effect for sitagliptin as expected, but did find a statistically significant incidence rate ratio (IRR >1.5) for primidone, an anticonvulsant not previously associated with liver toxicity. The study authors speculated that the false positive finding might have occurred because of concomitant therapy drugs, which were not considered in the OMOP analysis plan methods. Finally, the systematic analysis of the negative controls showed substantial bias, with results showing approximately 50 % increased risk of the target drugs for events where no risk should have been detected under the experiment assumptions.

This variability means future observational studies in similar data are highly likely to have inconsistent results and be hard to replicate. Relative risk estimates could be substantially biased in either direction. The sensitivity to method also means it would be simple to tinker with post hoc criteria, event definitions, and data selection to produce a null finding regardless of the true underlying risk. It also means that future observational studies sponsored by entities with a financial interest in the outcome must be examined and interpreted with additional care and skepticism.

4 European PROTECT Project

Like OMOP in the US, the European counterpart PROTECT was a partnership between the regulator, the EMA, pharmaceutical companies, and academic institutions. However, while OMOP was a large, tightly integrated scientific research program, PROTECT was more diverse. Its seven work programs [14] included objectives such as

pharmacovigilance training, better methods of communicating risks and benefits, and exploration of new methods of collecting adverse event information from consumers.

However, a key work program—framework for pharmacoepidemiology studies (WP2)—was structurally similar to OMOP. The published project design outlined five drug–adverse event pairs (for example, anticonvulsants and suicidal behaviors) [15]. The WP2 design called for the drug–adverse event pairs to be studied in six different European electronic record databases with common protocols that would evaluate different methods such as cohort, case control, and case crossover designs (similar to OMOP's self-controlled methods).

As far as can be determined, the grand design of WP2 could not be completed; at project end in February 2015 no published studies could be identified that met the published project design criteria. A survey of the 20 WP2 peer-reviewed publications shows that the investigators completed much more limited studies that did not evaluate drug–event relationships using different methods across the six databases. For example, the analysis of a possible relationship of anticonvulsant drugs and suicidal behaviors was not published—only a drug exposure study without the health outcome event [16]. Similarly, antidepressant use and hip fracture risk was explored only in a literature review [17]. A second published study in the WP2 program examined femur/hip fracture rates using European electronic health record databases but did not report on any possible association with drug treatment [18].

Nevertheless, the WP2 program research provided valuable insights into the obstacles to using electronic health data for postmarket surveillance. Ruigómez and colleagues evaluated the endpoint of acute liver injury (also an OMOP endpoint) in two primary care databases in Spain and the UK [19]. A manual review validated only 15 % of the cases in the Spanish database and 58.6 % in the widely used British Clinical Practice Research Datalink (CPRD). Furthermore, the incidence rates in the two populations varied threefold, a difference unlikely to reflect real differences of this magnitude in the health status of the two populations.

5 The Sentinel Initiative

The FDA's Sentinel electronic health data system for postmarket surveillance was first outlined in 2007 as an explicit mandate from the US Congress, even though no pilot of comparable size existed, no authoritative feasibility studies were on record, and nothing quite like it existed elsewhere. In addition, no funds to construct it were provided. Nevertheless, the law mandated an 'active postmarket risk identification system' that would include 25 million patients by 2010 and 100 million patients (or approximately one-third of the US population) by 2012 [20]. However, the law noted that which patient electronic health records could be obtained, how the data would be validated, and what analytical methods might be appropriate had not yet been determined. Nevertheless, despite these uncertainties, the formal start of Sentinel[1] in January 2009 did not lack for claims that a new era of postmarket surveillance had already arrived:

"This is a very important step we are taking today" said Michael Leavitt, then the Secretary of Health and Human Services and, at the time, the highest ranking health official in the federal government. "We are moving from reactive dependence on voluntary reporting of safety concerns—to proactive surveillance of medical products on the market. The result will be much improved safety protections for all Americans" [21].

By January 2015, Sentinel had become a formal reality as the FDA announced a transition from pilot project status to a completed system [22]. It claimed to include 178 million members. However, it had only 48 million individuals currently enrolled in 18 data partner health plans, and 35 million patients with more than 3 years of data [23].

5.1 Sentinel Design Features

The strengths and weaknesses of Sentinel revolve around several key design features that were established early and resulted from organizational and political obstacles as much as scientific considerations.

US medical care is provided and paid for in a heterogeneous collection of widely differing systems: single-payer systems serve three large populations—the elderly, military veterans, and active duty military. The 50 states run roughly similar, but independent, Medicaid programs to serve the poor. The largest fraction of the population, employed adults and their families, utilize private insurance carriers, which range from large insurers with tens of millions of covered patients to relatively small companies. In turn, the insurance carriers compensate thousands of different providers, varying widely both in size and care delivery structure.

These realities led the architects of Sentinel to collect primarily insurance claims or administrative data rather than the more extensive but diverse health records. Because of privacy concerns, the central coordinating center did not directly collect any patient-level data. Instead it created a common data model that would define information elements that would work across the 18 different data systems, and left it to each data partner to adapt their own records to

[1] Until 2015 the project was usually called 'Mini-Sentinel' and was regarded as a pilot.

the central model. To run an observational study, the coordinating center tested and developed SAS program code on the common data model, then distributed it to the partners to run. For much of its first 6 years, Sentinel was a work in progress, and the FDA reported spending US$150 million in external contracts between October 2009 and 2014.

5.2 Sentinel Results

Six years after the building of Sentinel commenced it appears that the system has not yet been the primary data source in identifying a single new drug risk that led to a significant regulatory action such as a drug withdrawal, boxed warning, restriction, or contraindication. The Mini-Sentinel website lists the following four drug safety actions [24] that involved its data.

- A comparison showed bleeding rates of dabigatran (PRADAXA) "did not appear to be higher" than warfarin in atrial fibrillation (a journal publication indicated Sentinel had captured only 16 cases of dabigatran gastrointestinal bleeds without age or gender information [25], and a later FDA electronic data study concluded that gastrointestinal bleeding rates for dabigatran were higher than for warfarin [26]).
- The FDA issued a warning of sprue-like enteropathy linked to the blood pressure medication olmesartan. However, Sentinel analysis was only one of four data sources, and the initial study results detected no increased risk.
- The Sentinel study of intussusception linked to two rotavirus vaccines for infants disclosed a small increased risk for one vaccine (RotaTeq), confirming a foreign study, but was inconclusive for the other (Rotarix).
- A Sentinel assessment of febrile seizures following immunization with the influenza vaccines in children under 5 years of age detected no statistically significant increased risk.

The reasons why Sentinel, heralded at creation as the centerpiece of a new era of postmarket surveillance, in fact produced such modest results did not appear to flow from problems unique to the project itself but rather resulted primarily from the current drawbacks of electronic health data. In Sects. 6–8 we examine the three critical problems seen in all three of the major research programs.

6 Terminology Problems

A significant and unavoidable limitation of searching for adverse drug events in electronic health data flows from the underlying terminologies through which patient encounters with the medical system are coded. Most electronic systems in the US rely on some variant of the World Health Organization's International Classification of Diseases, Ninth Revision (ICD-9) codes. One primary purpose of the terminology is to describe hospitalizations in sufficient clinical detail to document the care provided and procedures performed, and thereby establish a basis for payment [27]. The same codes are also used optionally in ambulatory care, where one assessment describes this situation as "the weakest link in the claims database" [27]. The terminology was never intended for the purpose of identifying adverse drug effects, and with just 28,000 terms to describe the universe of disease entities, it is not granular. This means any observational study based on ICD-9 codes begins with an elaborate hunt for relevant ICD-9 codes.

The first limitation is structural. Many kinds of drug adverse effects are seldom recorded as medical encounters, not only psychiatric side effects such as sexual dysfunction, suicide and aggression but also drug discontinuations because of adverse effects.

Even if a medical encounter generates an ICD-9 code indicating a possible drug effect (for example, dystonia), the terminology system allows only an optional additional E-code to identify the class of drugs that might be involved, and lacks codes for specific drugs. The number of ICD-9 codes generated by each encounter is variable, and may be influenced by the insurer's reimbursement policies [28]. In addition, ICD-9 codes are hierarchical, with the first three digits indicating a broad 'chapter' of disease entities, and additional digits providing more specificity. In some systems, or study algorithms, only the first three digits are used, making the system even less granular and less specific.

Several factors also can induce systematic bias. In the 1980s and 1990s many papers were written about 'upcoding' software that picked out ICD-9 codes that would trigger more generous hospital care payments [29]. Safety initiatives can also distort the data. Klompas [30] noted that safety initiatives to reduce ventilator-associated pneumonia spawned a substantial increase in diagnoses but no better outcomes, a paradox he attributed to increased but inaccurate diagnosis of the disorder.

To complicate matters, the ICD-9 codes may be implemented differently in different large claims databases. Madigan and colleagues [13] provided a graphic portrait of the problem as part of OMOP. The investigators took two widely used analytic methods (new user cohort and self-controlled case series) and evaluated the heterogeneity across the six large health claims databases in the project. They reported 30/106 (28 %) event–outcome pairs had statistically significant findings that were both positive and negative drug effects. In addition, 69/106 (65 %) had a range of point estimates in which the highest was more

than double the effect seen in the lowest. "Our findings suggest that 20–40 % of observational database studies can swing from statistically significant in 1 direction to statistically significant in the opposite direction depending on the choice of database…" the investigators concluded. These findings raise significant but unanswered questions about variability in Sentinel's 18 different databases.

Some limitations of ICD-9 codes cannot be remedied. Elements of a health encounter that were miscoded to enhance reimbursement, or just not accurately observed, or not recorded at all, cannot be repaired. However, other limitations, such as code selection for identifying the health outcome, can be addressed in validation studies. The US medical world is now transitioning from ICD-9 to the tenth revision (ICD-10), which is more granular and may eventually improve the accuracy and consistency of coded representations of health encounters.

While ICD-9 codes were central to the US studies, the European PROTECT project encountered the problem that different countries used substantially different terminologies. In a PROTECT study of the incidence of hip fractures [18], investigators needed widely varying event definitions. Two countries were using ICD-10 codes and required nine different codes to identify hip fracture events; two databases used International Classification of Primary Care (ICPC-2) using just one code; and two databases used the UK's Read terminology and had to specify 110 codes. Not only did the coding granularity vary between 1 and 110 event terms but the age-adjusted incidence rates varied more than twofold.

7 Data Validation

Given that these electronic data analysis programs were expected to produce drug safety findings robust and credible enough to support regulatory action, validation of the underlying electronic health data was essential. However, validation studies are expensive, typically do not produce exciting new scientific knowledge, and are labor intensive. The classic validation study method (more than 70 have been performed in the last 4 decades [31]) was uncomplicated: for any outcome event definition (e.g. upper gastrointestinal bleeding) the investigators' selection of ICD-9 codes was evaluated by checking against a modest sample of the available medical records, many of which proved to be unavailable or incomplete. A single positive predictive value (PPV) was then calculated, comparing a percentage of confirmed events with the total selected by the computer codes. One critical limitation of practically all such validation studies was that they were capable of assessing only one parameter of case identification: specificity, i.e. whether the medical records confirmed the outcome event identified through ICD-9 codes. Sensitivity—whether a few or many events were occurring but not reflected in the selected ICD-9 codes—has rarely been studied. Therefore, negative predictive values were generally unavailable to contribute an essential perspective on whether a group of ICD-9 codes was suitable for identifying the health outcome of interest.

The OMOP investigators did not conduct any validation studies in the six databases selected for their experiments; however, they did investigate the prevalence of various health outcomes depending on different event definitions, disclosing another source of result variability [32]. The published PROTECT work product included two event definition studies, which, as noted above, also revealed wide variability.

Validation was important to the Sentinel project because the FDA was constructing a national safety surveillance system for present and future use. Sentinel validation had two phases. The primary validation efforts consisted of literature searches for 20 health outcomes for drugs [33], and an additional 10 outcomes for vaccine injury surveillance [34]. The basic results showed the same problems seen with other aspects of electronic health records—large amounts of variability. Published studies varied widely in how much validation was performed, and how completely validation was reported. Underlying databases varied widely in population, size, and purpose, and the studies had been performed over 30 years' time. The reported PPVs were likely underestimated since most studies disregarded unavailable or incomplete medical records rather than counting them as unconfirmed observations as they would be in an intent-to-treat analysis in clinical trials.

In addition, Sentinel conducted direct medical record validity studies examining four health outcomes using hospitalization ICD-9 codes: acute kidney injury, severe acute liver injury, anaphylaxis, and acute myocardial infarction [35]. The best algorithms produced medical record-confirmed cases with PPVs that ranged from a low of 24.7 % for severe acute liver injury to 88 % for acute myocardial infarction. The small number of cases for each health outcome (range 129–143) did not permit measurement of differences among the 18 data partners.

In some instances, the ICD-9 codes selected to identify health outcomes could be a notable failure. A published study of suicides captured in administrative data had a PPV of only 14 %. In the Sentinel chart review, of the 56 cases of acute liver failure (a subset of the severe liver injury cases) just one was confirmed. Even anaphylaxis, a clearly defined medical emergency, produced a PPV of only 63 % in Sentinel partner data. Liver failure, although an important adverse drug event, produced consistently low PPVs across all three of the seminal research initiatives. The highest PPVs were generally seen for acute myocardial

infarction, often over 85 %. However, the high population prevalence, likelihood of comorbidities, numerous concomitant medications, and multiple causes render this health outcome challenging for identifying a single drug suspect.

The validity studies also demonstrate the variability and limitations of ICD-9 codes for correctly identifying health outcomes of interest, even when limited to fairly clear acute events that result in hospitalization. There was little or no recent validation data available for the majority of adverse drug events that did not usually result in hospitalization, e.g. tardive dyskinesia, hypertension, tachycardia, tics, weight gain, weight loss, diabetes, hypoglycemia, neutropenia, pruritus, impaired sexual function, cognitive impairment, cataracts, diarrhea, constipation, anxiety, insomnia, skin cancers, hyperuricemia, myopathy, and venous thromboembolism.

8 Statistical Methods

While statistical methods involved substantial questions about how best to analyze large health datasets, statistical methods were relatively easy to study, unlike expensive and difficult validation studies. Investigators sometimes used simulated data which permitted a pristine focus on method at the cost of uncertainty about whether the simulation reflected the realities of underlying health data. In addition, multiple methods could be applied to the same dataset with the only additional cost being a modest amount of programming in a statistical package. As a result, extensive studies were conducted and published. Methodology lay at the heart of the entire OMOP experiment and produced detailed assessments of seven commonly used statistical methods. Among the 20 publications to date from the European PROTECT WP2 effort, eight were diverse methods studies. Sentinel commissioned an early methods assessment, which examined the strengths and weaknesses of multiple disproportionality methods and sequential monitoring. Although a critical review of the results of these method studies would fill an entire book, these statistical issues have an impact on the credibility and value of existing and future studies.

8.1 Statistical Significance

The statistical significance p value is a standard parameter in frequentist statistics to characterize whether a finding is robust, and confidence intervals describe the uncertainty around the point estimate. However, for more prevalent outcomes and databases covering millions of patients, most results are likely to be statistically significant and the confidence intervals narrow. This proved true among the negative controls in OMOP, where scores of method–outcome pairs showed a statistically significant drug risk where none was expected. Much smaller scale but more expensive clinical trials are powered to be as small as possible while still being 80 % certain of detecting the hypothesized treatment effect. In this case, statistical significance and confidence intervals are parameters of central importance for interpreting clinical trial results. However, the same statistical significance measures in large electronic health data systems reveal little of interest, especially in the largest datasets and more prevalent health outcomes.

8.2 Failure to Disprove the Null Hypothesis

The opposite result—lack of statistical significance in comparisons between exposed and unexposed patients in electronic health data studies—creates even greater problems in interpreting the experiment. If the exposed patient population was relatively small, if the health outcome was rare, if event capture was weak, or if investigators used ICD-9 codes of narrow scope to increase specificity, then the number of index events could still be quite small, a few dozen or fewer. For example, the Sentinel study of dabigatran and warfarin in atrial fibrillation detected only 16 gastrointestinal hemorrhage events for dabigatran [25]; an FDA health record assessment of psychiatric hospitalizations for recent users of varenicline reported only 20 index events [36]. Such studies that fail to disprove the null hypothesis are also common in the literature. In a bibliography of observational studies published in 2013 from the British CPRD were 13 studies with no statistically significant findings, including assessments of acetaminophen and high blood pressure, metformin and lung cancer, opioids and type 2 diabetes, and or list at and acute liver injury [37]. In the Sentinel and CPRD cases, the data were insufficient to distinguish between an assurance of safety and an unsolved problem of type II error—design flaws or event ascertainment limitations that rendered the study incapable of detecting a difference if one existed.

An unresolved scientific issue of central importance is how to interpret a failure to disprove the null hypothesis in large electronic health data studies. A similar problem arises in randomized clinical trials that compare a new drug with an active control. When is the new treatment equivalent to the existing drug? In this instance, regulators require specific statistical design features that call for a larger sample size and additional parameters to build credible scientific evidence that the two treatments are equivalent. No such standards exist for observational studies in electronic health records. The experiments previously described here suggest that a null finding has a high probability of being a type II statistical error from the many

causes identified. Until clear and robust standards are devised to assess drug risk observational studies that do not detect statistically significant differences, little scientific weight should be attached to such studies.

8.3 Channeling and Confounding

A fundamental problem in observational studies for drug risk assessment is that patients exposed to the target drug are likely to be different from practically any comparison population selected. Patients prescribed an additional oral medication for type II diabetes are likely to be different from patients with similar glycemic control who were not prescribed a second or third agent. Data from the British CPRD showed a protective effect of naproxen on the risk of upper gastrointestinal bleeding—the opposite of the expected effect—apparently because of channeling, i.e. the physicians' deliberate selection of low-risk patients for naproxen [38]. One standard statistical approach to confounding is propensity score matching for which multiple methods exist. Uncertainties regarding propensity score matches involve several central questions, including whether propensity score matches exaggerate unmeasured confounders, thus increasing bias [39], the optimal method for achieving balance [40], and the adequacy of published reports in describing the success of the propensity score match [41]. In the OMOP results, case control and new user cohorts using propensity matches did not perform as well as self-controlled methods; however, because of limitations of underlying claims data, the propensity score match was limited to age and gender variables. The complex statistical method issues in propensity score matching are not unique to applications for postmarket surveillance; however, they add still another source of variability to any results apparently achieved.

9 Conclusions and Comments

There is no credible evidence that electronic health data today has the capacity to provide robust, reliable 'active surveillance', meaning identifying new drug risks not previously identified through other means. The results thus far dramatize the difficulties in confirming known adverse effects found using other methods.

The high levels of variability in almost every parameter render findings difficult to replicate and vulnerable to substantial bias, either as an accident of data and method selection or through intentional manipulation of study criteria.

At present, few studies have been conducted to assess the likelihood that risk assessments based on electronic health data systematically underestimate the adverse effects of drugs. Unless great caution is used in interpreting studies that do not detect a drug effect, society is at substantial risk that evidence of important drug harms may be masked, potentially blinding us to safety concerns that could affect millions of patients.

Nevertheless, electronic health data for drug safety risk assessment has substantial growth potential. The first priority is to address shortcomings in reliability, reproducibility, and statistical standards. This means additional validity studies, better understanding and disclosure of the kinds of adverse effects that will be captured poorly, and careful attention to improving the consistency and accuracy of the underlying electronic health data. Key findings from studies in one electronic health database should be reproducible and consistent with results from other datasets. Such reproducible findings should be compared with risk estimates from clinical trials and other sources. Future analysis needs to refine and broaden what has been learned about the preferred statistical methods, and new standards are needed to address the issue of interpreting studies that appear to detect no risk.

Substantial additional financial resources are also needed. When just one mid-sized clinical trial can cost tens of millions of dollars, the idea that clinical safety data of similar quality can be obtained from millions of electronic health records at minimal cost is naïve at best. Because it is a relatively complete system with an experienced coordinating center, the FDA's Sentinel, in particular, offers the opportunity for future improvement. Key software routines and data standardization, once completed and tested, can be reused at low cost and might meet the need for rapid results.

Finally, it is important for regulators and the medical community to understand that a critical element of the new paradigm—rapid and intensive drug surveillance through electronic health data—is nowhere near at hand. It does not now provide a viable safety net to counterbalance 'innovation promoting' drug approval policies that are reducing the number, size, rigor, and duration of randomized clinical trials.

Compliance with Ethical Standards *Funding* No sources of funding were used to assist in the preparation of this study.

Conflicts of interest Thomas J. Moore and Curt D. Furberg declare that they have no conflicts of interest relevant to the content of this manuscript.

Ethical approval This analysis relies exclusively on publicly available data and is therefore exempt from Institutional Review Board review requirements.

Acknowledgments The authors would like to thank Donald R. Mattison, McLaughlin Center for Population Health Risk Assessment, University of Ottawa, Ottawa, ON, Canada, for his comments and review of an earlier version of this article.

References

1. Hidalgo-Simon A, Arlett P. Pharmacovigilance in Europe: direction of travel in a changing environment. Expert Rev Clin Pharmacol. 2012;5:485–8.
2. Lumpkin MM, Eichler H-G, Breckenridge A, Hamburg MA, Lönngren T, Woods K. Advancing the science of medicines regulation: the role of the 21st-century medicines regulator. Clin Pharmacol Ther. 2012;92:486–93.
3. Baciu A, Stratton KR, Burke S. The future of drug safety: promoting and protecting the health of the public. Washington, DC: National Academies Press; 2007. Available at: http://www.nap.edu/catalog/11750/the-future-of-drug-safety-promoting-and-protecting-the-health. Accessed 12 Mar 2015.
4. Office of the Commissioner. FDA strategic priorities: 2014–2018. Silver Spring (MD): Food and Drug Administration; 2014. Available at: http://www.fda.gov/AboutFDA/ReportsManualsForms/Reports/ucm227527.htm. Accessed 26 May 2015.
5. Pharmacoepidemiological research on outcomes of therapeutics by a European consortium. PROTECT Home. 2014. Available at: http://www.imi-protect.eu/. Accessed 21 Jan 2015.
6. Observational Medical Outcomes Partnership (OMOP). Foundation for the National Institutes of Health. 2013. Available at: http://www.fnih.org/work/past-programs/omop. Accessed 26 Feb 2015.
7. Robb MA, Racoosin JA, Sherman RE, Gross TP, Ball R, Reichman ME, et al. The US Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. Pharmacoepidemiol Drug Saf. 2012;21:9–11.
8. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010;153:600–6.
9. Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, et al. A comparison of the empirical performance of methods for a risk identification system. Drug Saf. 2013;36(Suppl 1):S143–58.
10. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf. 2013;36(Suppl 1):S33–47.
11. DuMouchel W, Ryan PB, Schuemie MJ, Madigan D. Evaluation of disproportionality safety signaling applied to healthcare databases. Drug Saf. 2013;36(Suppl 1):S123–32.
12. Ryan PB, Schuemie MJ, Madigan D. Empirical performance of a self-controlled cohort method: lessons for developing a risk identification and analysis system. Drug Saf. 2013;36(Suppl 1):S95–106.
13. Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, et al. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol. 2013;178:645–51.
14. PROTECT Work Programme. PROTECT: Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium. 2014. Available at: http://www.imi-protect.eu/workprogram.shtml. Accessed 12 Mar 2015.
15. Abbing-Karahagopian V, Kurz X, de Vries F, van Staa TP, Alvarez Y, Hesse U, et al. Bridging differences in outcomes of pharmacoepidemiological studies: design and first results of the PROTECT project. Curr Clin Pharmacol. 2014;9:130–8.
16. De Groot MCH, Schuerch M, de Vries F, Hesse U, Oliva B, Gil M, et al. Antiepileptic drug use in seven electronic health record databases in Europe: a methodologic comparison. Epilepsia. 2014;55:666–73.
17. Prieto-Alhambra D, Petri H, Goldenberg JSB, Khong TP, Klungel OH, Robinson NJ, et al. Excess risk of hip fractures attributable to the use of antidepressants in five European countries and the USA. Osteoporos Int. 2014;25:847–55.
18. Requena G, Abbing-Karahagopian V, Huerta C, De Bruin ML, Alvarez Y, Miret M, et al. Incidence rates and trends of hip/femur fractures in five European countries: comparison using e-healthcare records databases. Calcif Tissue Int. 2014;94:580–9.
19. Ruigómez A, Brauer R, Rodríguez LAG, Huerta C, Requena G, Gil M, et al. Ascertainment of acute liver injury in two European primary care databases. Eur J Clin Pharmacol. 2014;70:1227–35.
20. 110th Congress Public Law 85, Section 905: active postmarket risk identification and analysis. Government Printing Office Public Laws. 2007. Available at: http://www.gpo.gov/fdsys/pkg/PLAW-110publ85/html/PLAW-110publ85.htm. Accessed 12 Mar 2015.
21. Leavitt M. Remarks as prepared at the Sentinel Press Conference. HHS.Govarchive. 2009. Available at: http://archive.hhs.gov/news/speech/2008/sp20080522a.html. Accessed 10 Feb 2015.
22. Woodcock J. Another important step in FDA's journey towards enhanced safety through full-scale "active surveillance". FDA Voice. 2014. Available at: http://blogs.fda.gov/fdavoice/index.php/2014/12/another-important-step-in-fdas-journey-towards-enhanced-safety-through-full-scale-active-surveillance/. Accessed 13 Jan 2015.
23. Mini-Sentinel. Mini-Sentinel distributed database "At A Glance". 2015. Available at: http://www.mini-sentinel.org/about_us/MSDD_At-a-Glance.aspx. Accessed 19 Mar 2015.
24. Mini-Sentinel. FDA safety communications. 2014. Available from: http://www.mini-sentinel.org/communications/fda_safety_communications/default.aspx. Accessed 9 Feb 2015.
25. Southworth MR, Reichman ME, Unger EF. Dabigatran and postmarketing reports of bleeding. N Engl J Med. 2013;368:1272–4.
26. Graham DJ, Reichman ME, Wernecke M, Zhang R, Southworth MR, Levenson M, et al. Cardiovascular, bleeding, and mortality risks in elderly Medicare patients treated with dabigatran or warfarin for nonvalvular atrial fibrillation. Circulation. 2015;131:157–64.
27. Strom BL. Data validity issues in using claims data. Pharmacoepidemiol Drug Saf. 2001;10:389–92.
28. Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. Pharmacoepidemiol Drug Saf. 2012;21:23–31.
29. Silverman E, Skinner J. Medicare upcoding and hospital ownership. J Health Econ. 2004;23:369–89.
30. Klompas M. The paradox of ventilator-associated pneumonia prevention measures. Crit Care. 2009;13:315.
31. West SL, Strom BL. Validity of pharmacoepidemiologic drug and diagnosis data. Pharmacoepidemiology. 4th ed. West Sussex: Wiley; 2005.
32. Reich CG, Ryan PB, Schuemie MJ. Alternative outcome definitions and their effect on the performance of methods for observational outcome studies. Drug Saf. 2013;36(Suppl 1):S181–93.
33. Carnahan RM, Moores KG, Perencevich EN. A systematic review of validated methods for identifying infection related to blood products, tissue grafts, or organ transplants using administrative data. Pharmacoepidemiol Drug Saf. 2012;21:213–21.
34. McPheeters ML, Sathe NA, Jerome RN, Carnahan RM. Methods for systematic reviews of administrative database studies capturing health outcomes of interest. Vaccine. 2013;31(Suppl 10):K2–6.
35. Mini-Sentinel. Validation of health outcomes| methods. Available from: http://www.mini-sentinel.org/methods/outcome_validation/default.aspx. Accessed 13 Mar 2015.
36. Meyer TE, Taylor LG, Xie S, Graham DJ, Mosholder AD, Williams JR, et al. Neuropsychiatric events in varenicline and nicotine replacement patch users in the Military Health System. Addiction. 2013;108:203–10.

37. Clinical Practice Research Datalink (CPRD). Available from: http://www.cprd.com/intro.asp. Accessed 10 Feb 2015.
38. Norén GN, Caster O, Juhlin K, Lindquist M. Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance. Drug Saf. 2014;37:655–9.
39. Brooks JM, Ohsfeldt RL. Squeezing the balloon: propensity scores and unmeasured covariate balance. Health Serv Res. 2013;48:1487–507.
40. Ali MS, Groenwold RHH, Pestman WR, Belitser SV, Roes KCB, Hoes AW, et al. Propensity score balance measures in pharmacoepidemiology: a simulation study. Pharmacoepidemiol Drug Saf. 2014;23:802–11.
41. Ali MS, Groenwold RHH, Belitser SV, Pestman WR, Hoes AW, Roes KCB, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. J Clin Epidemiol. 2015;68:112–21.