American Journal of Epidemiology

Vol. 178, No. 4
DOI: 10.1093/aje/kwt010
Advance Access publication:
May 5, 2013

Practice of Epidemiology

Evaluating the Impact of Database Heterogeneity on Observational Study Results

David Madigan*, Patrick B. Ryan, Martijn Schuemie, Paul E. Stang, J. Marc Overhage, Abraham G. Hartzema, Marc A. Suchard, William DuMouchel, and Jesse A. Berlin

* Correspondence to Dr. David Madigan, Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027 (e-mail: david.madigan@columbia.edu).

Initially submitted November 11, 2012; accepted for publication January 17, 2013.

Clinical studies that use observational databases to evaluate the effects of medical products have become commonplace. Such studies begin by selecting a particular database, a decision that published papers invariably report but do not discuss. Studies of the same issue in different databases, however, can and do generate different results, sometimes with strikingly different clinical implications. In this paper, we systematically study heterogeneity among databases, holding other study methods constant, by exploring relative risk estimates for 53 drug-outcome pairs and 2 widely used study designs (cohort studies and self-controlled case series) across 10 observational databases. When holding the study design constant, our analysis shows that estimated relative risks range from a statistically significant decreased risk to a statistically significant increased risk in 11 of 53 (21%) of drug-outcome pairs that use a cohort design and 19 of 53 (36%) of drug-outcome pairs that use a self-controlled case series design. This exceeds the proportion of pairs that were consistent across databases in both direction and statistical significance, which was 9 of 53 (17%) for cohort studies and 5 of 53 (9%) for self-controlled case series. Our findings show that clinical studies that use observational databases can be sensitive to the choice of database. More attention is needed to consider how the choice of data source may be affecting results.

database; heterogeneity; methods; population characteristics; reproducibility of results; surveillance

The increasing use of large-scale observational clinical databases underlies the recent rapid growth in the number of epidemiologic database studies. Such studies seek to use administrative claims data or electronic health records to address important questions about the effects of medical products by using observational study designs. Many potential biases and sources of variability threaten the validity of such studies, and a substantial literature documents these concerns (1–3). Although published studies typically discuss various limitations, such studies rarely discuss alternative databases that could have been used and how the choice of database might have affected results. Indeed, most reports simply identify the data source and provide no discussion about the process used to select it. The literature provides several important examples wherein different studies that used different data sources arrived at contradictory conclusions (4–7). Recent meta-analyses of observational studies have shown that individual studies of the same drug effect yielded conflicting results ranging from statistically significant decreased risk to statistically significant increased risk. Specific examples of meta-analyses wherein individual studies provide conflicting results include a meta-analysis (8) that considered the association between oral contraceptives and endometriosis, a meta-analysis (9) of the effects of proton pump inhibitors on mortality in patients receiving clopidogrel, and a meta-analysis (10) of the association between probiotics and sepsis in neonates. The literature on the cardiovascular risks associated with nonsteroidal antiinflammatory drug use is especially replete with conflicting observational studies (11). It is possible that a number of analyses that fail to show significant differences or are nonconfirmatory are never published, thus underestimating the true heterogeneity (12).

In these examples, multiple studies typically differ not only in their choice of data source but also in the designs and analytical methods applied to these data sources, so determining the contribution of each factor to heterogeneity can be challenging.

Table 1. Specific Databases Included in the Observational Medical Outcomes Partnership

Database Source	Database Description	Population Size, millions
General Electric Healthcare (Wauwatosa, Wisconsin)	Derived from data pooled from providers who use General Electric Centricity Office (an ambulatory electronic health record) into a data warehouse in a Health Insurance Portability and Accountability Act–compliant manner.	11.2
Truven Health Analytics, Inc. (Ann Arbor, Michigan)	MarketScan Lab Database represents a privately insured population with administrative claims from inpatient, outpatient, and pharmacy services supplemented by laboratory results.	1.5
	MarketScan Medicaid Multi-State Database contains administrative claims data for Medicaid enrollees from multiple states.	11.1
	MarketScan Medicare Supplemental and Coordination of Benefits Database captures administrative claims for retirees with Medicare supplemental insurance paid by employers, including services provided under Medicare-covered payment, employer-paid portion, and any out-of-pocket expenses.	4.4
	MarketScan Commercial Claims and Encounters Database represents a privately insured population and captures administrative claims with patient-level deidentified data from inpatient and outpatient visits and pharmacy claims of multiple insurance plans.	58
Humana, Inc. (Miramar, Florida)	Contains medical (inpatient, outpatient, and emergency room), pharmacy, and laboratory data (including test results) from Humana's administrative claims database of medical members.	6.5
Partners HealthCare (Boston, Massachusetts)	Includes data from the Partners HealthCare clinical transaction-based data repository as well as inpatient and outpatient billing feeds being collected in the research patient data registry, an analytical-structured database.	3
Regenstrief Institute, Indiana Network for Patient Care (Indianapolis, Indiana)	Contains population-based, longitudinal, and structured coded and text data captured from hospitals, physician practices, public health departments, laboratories, radiology centers, pharmacies, pharmacy benefit managers, and payers.	2
SDI Health, LLC (Plymouth Meeting, Pennsylvania)	Contains Health Insurance Portability and Accountability Act–compliant, deidentified, and encrypted patient-level data from hospitals, clinics, physician offices, and retail and specialty pharmacies from all 50 US states.	90
Department of Veterans Affairs (Washington, DC)	MedSAFE database includes data for US veterans from prescription dispensing records and electronic health records across facilities within the Veterans Affairs health care system as part of the Veterans Affairs Center for Medication Safety.	2

MATERIALS AND METHODS

In the present paper, we sought to isolate the effect of data source by holding all other aspects of the study design constant. We considered 53 drug-outcome pairs (Web Figures 1 and 2, available at http://aje.oxfordjournals.org/) representing a range of typical epidemiologic scenarios and, for each pair, applied 2 different study designs with the exact same implementation to 10 different databases. Note that the purpose of using 2 study designs is to compare the 2 methods only with respect to the variability in results among databases.

We previously conducted a study (13) to characterize the performance of epidemiologic designs applied in an automated fashion to large-scale observational health-care databases in predicting drug safety causal effects (http://omop.fnih.org). Specifically, our Observational Medical Outcomes Partnership project established a network of 10 data sources capturing the health-care experience of more than 130 million patients (Table 1). The data network included administrative claims data (from SDI Health, LLC, Plymouth Meeting, Pennsylvania; Humana, Inc., Miramar, Florida; and 4 MarketScan Research databases from Truven Health Analytics, Inc., Cambridge, Massachusetts, reflecting commercial claims with and without laboratory records, Medicare supplemental populations, and multistate Medicaid populations) and electronic health records (from Regenstrief Institute, Indiana Network for Patient Care, Indianapolis, Indiana; Partners HealthCare, Boston, Massachusetts; Centricity, General Electric Healthcare, Wauwatosa, Wisconsin; and the US Department of Veterans Affairs, Center for Medication Safety/Outcomes Research, Washington, DC). We transformed all of these data sets to a common data model, in which data about drug exposure and condition occurrence were structured in a consistent fashion and defined by using the same controlled terminologies, to facilitate subsequent analysis (14, 15).

A total of 13 different analytical methods were implemented during the Observational Medical Outcomes Partnership project. To simplify the presentation in the present study, we selected 2 widely used epidemiologic designs, namely a new-user cohort design with propensity score adjustment and a self-controlled case series design. The cohort design assembles a cohort of new users of the target drug and compares the rate of occurrence of the target outcome in this cohort with the same rate in a cohort of new

users of a comparison drug by using the propensity score approach to account for other differences between the cohorts. The self-controlled case series design estimates the association between a transient exposure and an adverse event by using only cases; no separate controls are required because each case acts as its own control. For all 53 drug-outcome pairs (representing 9 "positive" and 44 "negative" controls) (Web Appendix 1), we applied the cohort method in a consistent and typical fashion, whereby we required an exposure-free observation period of at least 180 days and then included outcomes that occurred within 30 days of the start of exposure. A maximum of 100 empirically derived covariates (16) were included in a logistic regression used to estimate a propensity score, which was then used to stratify the population into 20 strata. The comparison drugs in each case were drugs with the same indication as for the target drug but were not in the same drug class. For the self-controlled case series method for all 53 pairs, we considered the first occurrence of each outcome, excluded outcomes occurring on the first day of any exposure period, and used a variance of 2 for the normal regularizing prior on the treatment effect. We executed the cohort method against all but 1 database (SDI Health), and the self-controlled case series method was run for all but 1 database (Partners HealthCare), so within-method comparisons include 9 databases. Complete descriptions, references, and source code for each method are available at http://omop.fnih.org/MethodsLibrary. For each combination of design and database, we generated relative risks and standard error estimates across the 53 drug-outcome test cases. We reported a result as statistically significant if the 2-sided P value was less than 0.05. We present the results separately for the 2 methods.

For each of 53 drug-outcome pairs and for each of 2 standard epidemiologic methods, we calculated an estimated relative risk and associated standard error in each of 9 databases. Our primary interest concerned the heterogeneity of these 9 estimates across the databases. As noted above, one could debate the specific choices of design, but there is no reason to suspect, a priori, that the design choices would affect different databases differently. To assess the heterogeneity, we considered 1) the I^2 heterogeneity statistic (i.e., the normalized Cochran's Q statistic) for each pair, summarized within each study design (following Higgins et al. (17), we used a threshold of 75% to identify "high" heterogeneity and 25% to identify "low" heterogeneity); 2) consistency of statistical significance; and 3) a graphical presentation of the results.

RESULTS

Table 2 shows the number of persons in each database for each outcome and drug of interest. For many drug-outcome pairs and for both methods we observed substantial heterogeneity across databases. The full data set containing all drug-outcome estimates across all sources for the 2 methods discussed here is available in Web Appendix 2. The estimates for all other methods and parameter settings are available in Web Appendix 3. The specific parameter settings are detailed in Web Appendix 4.

New user cohort design

For the new user cohort design across all 53 pairs, we observed 23 drug-outcome pairs (43%) with $I^2 \geq 75\%$ and 13 drug-outcome pairs (25%) with $I^2 \leq 25\%$. Eleven pairs (21%) had at least 1 source with a significant positive effect and at least 1 source with a significant negative effect. For 25 of the pairs (47%), the range in point estimates exceeded 0.693 on a natural log scale (equivalent to 1 database having an effect size estimate at least twice that of another). In 23 cases, all estimates had point estimates that were directionally consistent (positive or negative), of which 9 cases were statistically significant in all sources.

Web Figure 1 shows the relative risk point estimates color-coded according to statistical significance. Most of the pairs that provided statistically contradictory evidence (i.e., were statistically significantly above 1.0 in 1 database and statistically significantly below 1.0 in another database) had combined estimated relative risks nearer to 1.0; equivalently, most of the pairs that did not exhibit such contradictions were close to either the top or the bottom (i.e., they had large relative risks either above or below 1.0). There was no apparent consistency as to which databases were significantly negative (suggesting decreased risk), nonsignificant, or significantly positive (suggesting increased risk). Each database was inconsistent with other databases on at least 1 drug-outcome pair.

Self-controlled case series

For the self-controlled case series design across all 53 pairs, we observed 37 drug-outcome pairs (70%) with $I^2 \geq 75\%$ and 7 pairs (13%) with $I^2 \leq 25\%$. Nineteen pairs (36%) had at least 1 source with significant positive effect and at least 1 source with significant negative effect. Forty-four pairs (83%) had a range in point estimates greater than 0.693 on the log scale (equivalent to 1 database having an effect size estimate at least twice that of another). In 18 cases, all databases had point estimates that were directionally consistent (positive or negative), of which 5 cases were statistically significant in all sources.

DISCUSSION

Many different potential biases and sources of variability can undermine the validity of epidemiologic analysis of observational databases. Even when holding data source constant, heterogeneity can persist, presumably because of observed and unobserved patient characteristics that vary across databases. For example, 2 recent studies (18, 19) investigated the association between bisphosphonate use and esophageal cancer by using the same data source (the United Kingdom General Practice Research Database) yet arrived at different conclusions. Our study, however, focused specifically on the database as a source of heterogeneity. 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, despite holding study design constant. We also found that almost all studies can be

Table 2. Number of Persons in Each Database for Each Outcome and Drug of Interest, Observational Medical Outcomes Partnership

Outcomes and Drugs of Interest	Database									
	MSLR[a]	MDCD[b]	MDCR[c]	CCAE[d]	GE[e]	Regenstrief[f]	Humana[g]	Partners HealthCare[h]	SDI Health[i]	VA MedSAFE[j]
Health outcome of interest										
Angioedema	3,931	19,667	14,844	96,703	9,306	477	10,969	6,966	121,116	10,376
Aplastic anemia	2,674	23,022	32,106	46,838	6,261	13	8,896	10,096	100,324	7,466
Bleeding	298,843	1,328,112	1,277,716	6,126,762	682,689	17,797	763,382	389,666	10,062,046	686,663
Hip fracture	2,206	66,046	126,360	49,600	13,491	6,749	29,870	16,988	448,320	14,632
Hospitalization	962,990	7,323,336	3,637,637	28,669,601	289,366	1,216,746	2,613,318	762,717	62,212,614	703,433
Liver failure (acute)	198,213	1,430,934	613,148	4,647,376	626,444	11,260	461,697	264,088	6,491,416	346,619
Mortality after myocardial infarction	406	19,002	29,668	12,674	7,321	48	22,696	37,624		21,860
Myocardial infarction (acute)	28,846	183,192	420,769	660,227	62,027	1,287	161,022	107,699	1,620,080	146,839
Renal failure (acute)	9,900	133,461	208,416	147,623	13,337	2,146	102,346	44,802	946,331	110,628
Upper gastrointestinal ulcer (requiring hospitalization)	36,298	204,344	310,948	717,064		2,237	113,623	21,637	1,042,263	38,236
Target drug of interest										
ACE inhibitor	108,869	614,703	1,669,766	3,062,264	1,361,068	126,940	2,292,064	172,363	9,693,462	1,896,816
Amphotericin B	64	381	326	1,743	222	62	162	6,071	2,809	46
Antibiotics	106,423	1,190,606	709,306	3,700,609	744,632	161,049	1,048,639	72,440	8,022,639	334,248
Antiepileptics	10,434	161,887	74,638	182,629	98,821	7,974	129,019	36,997	669,119	81,469
Benzodiazepines	129,742	879,378	1,034,022	3,826,110	1,038,394	140,683	1,123,034	269,209	10,666,276	633,267
β blockers	82,747	498,144	1,389,476	2,331,296	1,139,129	101,974	1,913,788	248,614	8,166,071	1,610,374
Bisphosphonates	16,909	72,068	413,216	427,966	246,091	19,446	473,346	32,634	1,786,837	118,621
Tricyclic antidepressants	30,676	320,962	268,908	889,212	361,764	34,927	362,226	63,991	2,398,909	276,640
Typical antipsychotics	14,946	174,663	161,627	322,722	94,186	14,611	196,660	78,873	1,216,213	83,412
Warfarin	14,164	144,440	616,291	373,664	314,634	22,309	679,680	89,203	2,404,321	286,988

Abbreviation: ACE, angiotensin-converting enzyme

[a] MarketScan Lab Database (Truven Health Analytics, Inc., Ann Arbor, Michigan) ($n = 1,466,617$).

[b] MarketScan Medicaid Multi-State Database (Truven Health Analytics, Inc., Ann Arbor, Michigan) ($n = 11,188,360$).

[c] MarketScan Medicare Supplemental and Coordination of Benefits Database (Truven Health Analytics, Inc., Ann Arbor, Michigan) ($n = 4,666,736$).

[d] MarketScan Commercial Claims and Encounters Database (Truven Health Analytics, Inc., Ann Arbor, Michigan) ($n = 69,836,290$).

[e] Centricity Medical Records Database (General Electric Healthcare, Wauwatosa, Wisconsin) ($n = 11,216,208$).

[f] Regenstrief Institute database (Indiana Network for Patient Care, Indianapolis, Indiana) ($n = 2,002,480$).

[g] Humana database (Humana, Inc., Miramar, Florida) ($n = 9,348,480$).

[h] Partners HealthCare database (Partners HealthCare, Boston, Massachusetts) ($n = 2,942,640$).

[i] SDI Health database (SDI Health, LLC, Plymouth Meeting, Pennsylvania) ($n = 90,864,482$).

[j] MedSAFE database (US Department of Veterans Affairs, Washington, DC) ($n = 3,201,630$).

statistically significant in some databases and statistically nonsignificant in others. Although variability in statistical power across the databases could explain this latter finding, it does not explain the former finding of contradictory statistical significance.

Of the 2 methods evaluated in our study, the cohort approach exhibited less among-database heterogeneity than did the self-controlled case series. In general, the self-controlled case series results in smaller standard errors and hence more statistically significant effect estimates; this may partly explain the difference.

Our analyses included outcomes that occurred within 30 days of the start of exposure. Although the choice of 30 days might not be optimal for every drug-exposure pair because it focuses on short-term effects, it is a commonly used "window" for ascertaining events. Further, for purposes of this paper, the point is that consistent application of this approach should still produce similar results across databases.

In the face of high heterogeneity, simply pooling data or performing a meta-analysis will generally not provide satisfactory outputs (17, 20, 21), not least because the final results will derive largely from the particular choice of databases included in the analysis and the relative weights associated with those databases in the meta-analysis. Random-effects meta-analyses allow explicitly for heterogeneity, but causal interpretation of random-effects meta-analytical outputs remains problematic (22). The Cochrane Handbook (http://www.cochrane-handbook.org/) points out, "a random-effects meta-analysis model does not 'take account' of the heterogeneity in the sense that it is no longer an issue." Systematic approaches to understanding and characterizing among-study heterogeneity represent an important part of the meta-analysis process (23), but identifying specific elements that explain variability across observational data can prove challenging. Certainly, deriving a composite estimate in the face of significant heterogeneity should be discouraged. Moreover, observing large heterogeneity should raise questions about the ability of observational data to address the clinical question at all. We recommend that all pooled estimates or meta-analyses should, at a minimum, present study-specific estimates (possibly with cross-database shrinkage) in addition to quantitative measures of heterogeneity.

The 10 databases that we included in this study differ in terms of covered populations, completeness of the data capture, patient susceptibility to adverse events, and the accuracy of the recorded information. We included data that arose from different components of the health-care system, such as administrative claims that arise in the reimbursement process between providers and payers and electronic health-care records that arise at the point of interaction between patients and providers. However, this is central to the point this study intends to make, because published studies in fact use a similar diversity of databases. The databases are not in the public domain, but researchers can access them through the Observational Medical Outcomes Partnership project.

Why do different databases show different biases? It is well understood that administrative claims and the electronic health records database originate from different data capture processes, neither of which has research as a primary intention. Claims data represent data elements captured as part of the reimbursement process; therefore, they reflect diagnosis codes captured to justify procedures, other services, and pharmacy dispensings that are processed as part of a prescription benefit (resulting in underrepresentation of over-the-counter drug exposures). Electronic health records data are captured to support clinical care; therefore, they reflect the information that providers require to perform services, which may not represent a complete medical history (24, 25). Beyond the data capture process, each database reflects a different source population with varied patient demographics, underlying disease severity, and length of longitudinal data capture. Patients may receive care from different health-care systems and over different periods of time, so differences in geographic and temporal quality of care could alter data captured for those individuals.

A better understanding of heterogeneity due to data source can also inform the appropriate application of distributed data networks, such as the US Centers for Disease Control and Prevention's Vaccine Safety Datalink (http://www.cdc.gov/vaccinesafety/activities/vsd.html) and the US Food and Drug Administration's Mini-Sentinel (http://mini-sentinel.org/). Such networks combine multiple disparate data sources, each contributing summary statistics, and have particular promise for the study of rare effects. Significant heterogeneity due to data source, however, may undermine the usefulness of such networks. As best practice, all network-based analysis should be required to present forest plots that show database-specific estimates as a standard output.

In practice, effect size estimates and associated confidence intervals are more useful than statistical significance. For simplicity, we have emphasized statistical significance in our presentation, but our basic point about heterogeneity due to database pertains either way.

We are unaware of any previous attempt to systematically characterize database heterogeneity in observational database studies. One limitation of our study is that the 53 drug-outcome pairs arose in the context of a drug safety study and, as such, primarily relate to safety rather than effectiveness. It is possible that effectiveness studies exhibit less heterogeneity. Our results pertain to observational studies and not randomized trials. Some authors (25, 26), however, have suggested that observational data based on records from daily practice are more germane to safety questions than are data from randomized trials, and certainly observational databases are routinely used to address both effectiveness and safety questions. We applied each method (i.e., cohort and self-controlled case series) in exactly the same way for all 53 pairs. It is conceivable that customizing the analytical approach to each pair could lead to greater consistency across databases. We have, in fact, conducted analyses across all drug-outcome pairs and all databases by using hundreds of different possible design choices (e.g., different washout periods, different surveillance windows) (data available at http://omop.fnih.org/OMOP2011Symposium) and failed to identify a set of choices that produced substantially less heterogeneity.

Note that in this study we have not attempted to fully characterize the differences in participants, exposures, and outcome measurements in the different databases and instead focused on quantifying the extent of the observed heterogeneity. Standard epidemiologic analytical methods such as the self-controlled case series and the new user cohort design that we have used in our analyses attempt to control for within-data-source confounding. Our analyses demonstrate that these standard approaches fail to account for the various characteristics that produce the between-source variation that our analysis highlights.

A reviewer of our manuscript suggested 2 strategies for placing the observed heterogeneity in context. The first strategy combines all the databases and then assesses heterogeneity across sampled subdatabases from the pooled databases; in essence, this provides a kind of null distribution against which to compare observed heterogeneity. The second strategy also pools the databases but considers heterogeneity across subgroups defined by, for example, age and sex. Neither strategy is possible in our distributed environment, but both represent interesting future lines of research in a centralized environment.

We note that in our analysis, each of the databases is quite large (millions of persons) with large numbers of exposures and outcomes. As such, pooling study results for the purposes of gaining precision is of secondary concern, and instead the value of disparate databases is producing a more accurate assessment of the effect size and a more comprehensive understanding of systematic error that may be influencing results.

We believe our findings have 2 immediate implications. First, when interpreting results from a single observational data source, more attention is needed to consider how the choice of data source may be affecting results. Second, where possible, studies should examine multiple sources to confirm that significant findings are consistently identified, or that results are at least consistent across databases. When interpreting results across multiple sources, it is important to characterize the observed heterogeneity and limit the use of composite estimates that could otherwise hide the uncertainty in effect estimates that is not driven by sampling variability. Forest plots provide especially useful insights into heterogeneity and should always be included. We believe our findings have relevance to authors of individual studies as well as to authors of review articles. Individual studies carried out in a single database should acknowledge that the study findings might be different in different databases. Review articles should note that component studies as well as meta-analytical findings raise the same concern.

ACKNOWLEDGMENTS

Author affiliations: Department of Statistics, Columbia University, New York, New York (David Madigan); Janssen Research and Development, LLC, Titusville, New Jersey (Patrick B. Ryan, Paul E. Stang, Jesse A. Berlin); Department of Medical Informatics, Erasmus University Medical Center Rotterdam, Rotterdam, the Netherlands (Martijn Schuemie); Regenstrief Institute, Center for Biomedical Informatics, and Indiana University School of Medicine, Indianapolis, Indiana (J. Marc Overhage); College of Pharmacy, University of Florida, Gainesville, Florida (Abraham G. Hartzema); Departments of Biomathematics and Human Genetics, David Geffen School of Medicine at the University of California, Los Angeles, Los Angeles, California (Marc A. Suchard); Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California (Marc A. Suchard); Oracle Health Sciences, Burlington, Massachusetts (William DuMouchel); and Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, Maryland (David Madigan, Patrick B. Ryan, Martijn Schuemie, Paul E. Stang, J. Marc Overhage, Abraham G. Hartzema, Marc A. Suchard, William DuMouchel, Jesse A. Berlin). At the time of this work, Abraham G. Hartzema was on sabbatical at the US Food and Drug Administration.

The Observational Medical Outcomes Partnership is funded by the Foundation for the National Institutes of Health through generous contributions from the following: Abbott Laboratories, Ltd.; Amgen, Inc.; AstraZeneca, PLC; Bayer Healthcare Pharmaceuticals, Inc.; Bristol-Myers Squibb Company; Eli Lilly & Co.; GlaxoSmithKline, PLC; Janssen Research and Development, LLC, Lundbeck, Inc.; Merck & Co., Inc.; Novartis Pharmaceuticals Corp.; Pfizer, Inc.; Pharmaceutical Research Manufacturers of America; F. Hoffmann-La Roche, Ltd.; Sanofi, SA; Schering-Plough Corp.; and Takeda Pharmaceutical Co., Ltd.

Although Drs. Ryan, Stang, and Berlin are employees of Janssen Research and Development, LLC, and Dr. Ryan is a past employee of GlaxoSmithKline, PLC, the authors declare no conflict of interest.

REFERENCES

1. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
2. Davey Smith G, Ebrahim S. Epidemiology—is it time to call it a day? *Int J Epidemiol*. 2001;30(1):1–11.
3. Mayes LC, Horwitz RI, Feinstein AR. A collection of 56 topics with contradictory results in case-control research. *Int J Epidemiol*. 1988;17(3):680–685.
4. Lockhart PB, Bolger AF, Papapanou PN, et al. Periodontal disease and atherosclerotic vascular disease: Does the evidence support an independent association? A scientific statement from the American Heart Association. *Circulation*. 2012; 125(20):2520–2544.
5. Dodd S. Debating the evidence: oral contraceptives containing drospirenone and risk of blood clots. *Curr Drug Saf*. 2011; 6(3):132–133.
6. Juni P, Nartey L, Reichenbach S, et al. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *Lancet*. 2004;364(9450):2021–2029.
7. Young SS, Karr A. Deming, data and observational studies. *Significance*. 2011;8(3):116–120.
8. Vercellini P, Eskenazi B, Consonni D, et al. Oral contraceptives and risk of endometriosis: a systematic review and meta-analysis. *Hum Reprod Update*. 2011;17(2):159–170.

9. Kwok CS, Loke YK. Meta-analysis: the effects of proton pump inhibitors on cardiovascular events and mortality in patients receiving clopidogrel. *Aliment Pharmacol Ther*. 2010;31(8):810–823.
10. Deshpande G, Rao S, Patole S, et al. Updated meta-analysis of probiotics for preventing necrotizing enterocolitis in preterm neonates. *Pediatrics*. 2010;125(5):921–930.
11. Hernandez-Diaz S, Varas-Lorenzo C, Garcia Rodriguez LA. Non-steroidal antiinflammatory drugs and the risk of acute myocardial infarction. *Basic Clin Pharmacol Toxicol*. 2006; 98(3):266–274.
12. Dwan K, Altman DG, Arnaiz JA, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS One*. 2008;3(8):e3081.
13. Ryan PB, Madigan D, Stang PE, et al. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat Med*. 2012;31(30):4401–4415.
14. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med*. 2010;153(9):600–606.
15. Overhage JM, Ryan PB, Reich CG, et al. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc*. 2012;19(1):54–60.
16. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4):512–522.
17. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414): 557–560.
18. Cardwell CR, Abnet CC, Cantwell MM, et al. Exposure to oral bisphosphonates and risk of esophageal cancer. *JAMA*. 2010;304(6):657–663.
19. Green J, Czanner G, Reeves G, et al. Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort. *BMJ*. 2010;341:c4444.
20. Maclure M. Demonstration of deductive meta-analysis: ethanol intake and risk of myocardial infarction. *Epidemiol Rev*. 1993;15(2):328–351.
21. Thompson SG, Pocock SJ. Can meta-analyses be trusted? *Lancet*. 1991;338(8775):1127–1130.
22. Greenland S. Can meta-analysis be salvaged? *Am J Epidemiol*. 1994;140(9):783–787.
23. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *JAMA*. 2000;283(15):2008–2012.
24. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323–337.
25. Hennessy S. Use of health care databases in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*. 2006;98(3):311–313.
26. Vandenbroucke JP. Why do the results of randomised and observational studies differ? *BMJ*. 2011;343:d7020.