Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies

Ther Adv Drug Saf

(2013) 4(2) 53–62

DOI: 10.1177/2042098613477445

David Madigan, Patrick B. Ryan and Martijn Schuemie

Abstract:

Background: Clinical studies that use observational databases, such as administrative claims and electronic health records, to evaluate the effects of medical products have become commonplace. These studies begin by selecting a particular study design, such as a case control, cohort, or self-controlled design, and different authors can and do choose different designs for the same clinical question. Furthermore, published papers invariably report the study design but do not discuss the rationale for the specific choice. Studies of the same clinical question with different designs, however, can generate different results, sometimes with strikingly different implications. Even within a specific study design, authors make many different analytic choices and these too can profoundly impact results. In this paper, we systematically study heterogeneity due to the type of study design and due to analytic choices within study design.

Methods and findings: We conducted our analysis in 10 observational healthcare databases but mostly present our results in the context of the GE Centricity EMR database, an electronic health record database containing data for 11.2 million lives. We considered the impact of three different study design choices on estimates of associations between bisphosphonates and four particular health outcomes for which there is no evidence of an association. We show that applying alternative study designs can yield discrepant results, in terms of direction and significance of association. We also highlight that while traditional univariate sensitivity analysis may not show substantial variation, systematic assessment of all analytical choices within a study design can yield inconsistent results ranging from statistically significant decreased risk to statistically significant increased risk. Our findings show that clinical studies using observational databases can be sensitive both to study design choices and to specific analytic choices within study design.

Conclusion: More attention is needed to consider how design choices may be impacting results and, when possible, investigators should examine a wide array of possible choices to confirm that significant findings are consistently identified.

Keywords: analysis, healthcare database, health outcomes, study design

Correspondence to:
David Madigan PhD
Professor and Chair, Department of Statistics, Columbia University, 1255 Amsterdam Ave., New York, NY 10027, USA
david.madigan@columbia.edu

Patrick B. Ryan
Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, MD and Janssen Research and Development LLC, Titusville, NJ, USA

Martijn Schuemie
Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, MD and Department of Medical Informatics, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands

Introduction

Observational studies using large-scale healthcare databases inform an increasing range of medical policy and practice decisions. Important data sources include administrative claims and electronic health record systems. Many potential biases and sources of variability threaten the validity of such studies and a substantial literature documents these concerns [Ioannidis, 2005; Smith and Ebrahim, 2001; Perrio *et al.* 2007; Mayes *et al.* 1988]. While authors adopt a variety of statistical and epidemiological approaches, analyses based on cohort, case control, and self-controlled designs dominate the literature. Indeed, most reports simply state the primary design choice, cohort, case control, or self-controlled, and provide no discussion about the process used to select it. The widely cited Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines for reporting of observational studies call for a clear statement of the study design but do not mention discussion of the rationale for the choice of study design [von Elm *et al.* 2008]. Similarly, within each of these choices, reports usually provide little or no justification for the many subsequent analytic choices, such as definition of time at risk, selection of covariates, and length of washout period.

Consider, for example, observational studies of the association between nonsteroidal anti-inflammatory drugs (NSAIDs) and myocardial infarction. Hernandez-Diaz and colleagues conducted a meta-analysis of observational studies as of 2006 [Hernandez-Diaz *et al.* 2006]. The authors included 16 studies in total, 4 cohort studies, 9 case control studies, and 3 nested case control studies. All three designs are represented in the subset of the studies that used administrative claims databases. Within each design, different studies made different analytic choices. For example, within the nested case control studies, the time-at-risk choices included on drug, on drug plus 30 days, on drug plus 90 days, and on drug plus 180 days.

Many observational studies have examined the association between thiazolidinediones and cardiovascular outcomes. Loke and colleagues considered 16 studies that compared rosiglitazone and pioglitazone with regard to the risk of myocardial infarction, 4 nested case control studies, and 12 cohort studies [Loke *et al.* 2011]. Again, within each study design, analytic choices varied across the studies. For example, one cohort study excluded patients who had received insulin within 365 days prior to exposure, another excluded patients currently using insulin, while others took no account of insulin use. No two studies adjusted for the same set of potential confounders. Among the cohort studies, some opted to evaluate all exposed patients while others focused on new users; among the new user designs, some chose an exposure washout period of 6 months while others required 12 months of observation prior to first exposure.

Studies of the same clinical question with different analytic approaches can generate different results, sometimes with strikingly different implications. Two recent studies concerning oral bisphosphonates and the risk of esophageal cancer illustrate the cause for concern. One study conducted a case control analysis in the UK General Practice Research Database (GPRD) and concluded the ‘risk of oesophageal cancer increased with 10 or more prescriptions for oral bisphosphonates and with prescriptions over about a five year period’ [Green *et al.* 2010]. The other study conducted a cohort analysis, also in the GPRD, and concluded, ‘the use of oral bisphosphonates was not significantly associated with incident esophageal or gastric cancer’ [Cardwell *et al.* 2010]. While the two studies differed in terms of the primary design choice, one using case control and the other using a cohort design, they agreed on some of the subsequent analytic choices. Both studies excluded patients aged 40 and under, and both studies adjusted for smoking, alcohol, and body mass index, but disagreed on others. One study adjusted for hormone therapy while the other did not, and one study considered patients with prescriptions between 1995 and 2005 whereas the other included 1996–2006. Dixon and Solomon explored the differences between these studies at some length [Dixon and Solomon, 2011]. De Vries and colleagues considered a different pair of GPRD studies that arrived at different conclusions [De Vries *et al.* 2006]. Both studies considered studies of fracture risk and statins, with one [Meier *et al.* 2000] showing a statistically significant benefit and the other [van Staa *et al.* 2001] showing a benefit, but falling short of statistical significance. Design choices that differed between these two studies included the age band used for matching cases and controls, the selection of potential confounders, the exclusion of high-risk patients, and different definitions for exposure time windows. In this paper we provide a systematic analysis of the impact of study design choices on observational results.

Methods

To study the impact of analytic design choices, such as those reflected in the GPRD studies, we conducted a series of analyses for a set of drug–outcome pairs across 10 databases, but mostly illustrate the results in the GE Centricity (GE) database, an ambulatory electronic health record database. The GE database contains data for 11.2 million lives. Men comprise 42% of the database, and the mean age is 39.6 years. The data represent 22.4 million patient years covering the time period 1996–2008. The GE MQIC (Medical

Quality Improvement Consortium) represents the group of providers who use the GE Centricity Electronic Medical Record and who contribute their data for secondary analytic use. The GE database reflects events in usual care, including patient problem lists, prescribing patterns and over-the-counter (OTC) use of medications, and other clinical observations as experienced in the ambulatory care setting. Because GE primarily reflects outpatient services, inpatient care and linkage across providers is under represented. A number of publications provide descriptions of various characteristics of the GE database [Brixner *et al.* 2007; Crawford *et al.* 2010; Gill and Chen, 2008].We transformed the dataset to the Observational Medical Outcomes Project (OMOP) common data model, in which data about drug exposure and condition occurrence were structured in a consistent fashion and defined using the same controlled terminologies to facilitate subsequent analysis (see http://omop.fnih.org) [Overhage *et al.* 2011; Stang *et al.* 2010]. We note that the mapping has the potential to lose important medical information [Reich *et al.* 2012]. We conducted identical analyses in nine other databases [Ryan *et al.* 2012], also mapped to the OMOP common data model [Overhage *et al.* 2011] that led to qualitatively similar conclusions and we make the corresponding data available online.

For this study, we selected three widely used epidemiologic designs: a new-user cohort design with propensity score adjustment, a case control design, and a self-controlled case series (SCCS) design. For illustration purposes we will use the same drug as in the motivating example, that is, oral bisphosphonates. We consider the association between oral bisphosphonates and four outcomes that are considered important for drug safety [Stang *et al.* 2011; Trifiro *et al.* 2009]: aplastic anemia, acute renal failure, acute myocardial infarction, and acute liver injury. We do not believe there is an association between these outcomes and bisphosphonates based on thorough reviews of product labels and published literature [Brixner *et al.* 2007]. Therefore, absent bias, and excepting analyses with active comparators that have a protective or harmful effect, most analyses should produce estimates that are consistent with no effect (e.g. relative risk confidence intervals that straddle one). We note that while the interpretation of our findings does depend on this assumption, our findings concerning heterogeneity do not.

To examine the impact of type of study design, we implemented cohort, case control, and self-controlled designs. We varied specific design features within each method, with the goal of including the kinds of analyses commonly seen in the literature. For each design we chose one particular set of design features to serve as the typical analysis for that design. Later in the paper we consider the impact of varying these within-design features. One typical application of the cohort method that we applied to all databases required an exposure-free observation period of at least 180 days and included outcomes that occurred during exposure or within 30 days after the end of exposure. A maximum of 100 covariates were included in a logistic regression used to estimate a propensity score, which was in turn used to stratify the population into five strata. We used a variable selection algorithm described by Schneeweiss and colleagues to select the covariates [Schneeweiss *et al.* 2009]. We estimated relative risk by comparing new users of bisphosphonates to patients exposed to an active comparator drug, raloxifene, a drug that shares the same primary indication as bisphosphonates (osteoporosis) but reflects a different mechanism of action. These design choices are similar to those chosen by Schneeweiss and colleagues in their exploration of NSAIDs and gastrointestinal bleeding [Schneeweiss *et al.* 2009]. For the typical case control design we chose four controls per case, each matched on age and sex, required a 183-day period of observation in the database prior to the index date, as well as a 30-day period of observation after the event (for cases) or index date (for controls). We included incident outcomes that occurred during exposure or within 30 days. For the typical SCCS method, we considered the first occurrence of each outcome, excluded outcomes occurring on the first day of any exposure period, and used a variance of 1 for the normal regularizing prior to the treatment effect. We counted ‘exposed’ outcomes that occurred during exposure or within 30 days. If possible, choices that span all three designs, such as time at risk, were held constant. Complete descriptions, references, and source code for each method are available at http://omop.fnih.org/MethodsLibrary. We report a result as statistically significant if the two-sided p value is smaller than 0.05.

We present four sets of results. First we present results for each of the four outcomes for each study design, cohort, case control, and SCCS.

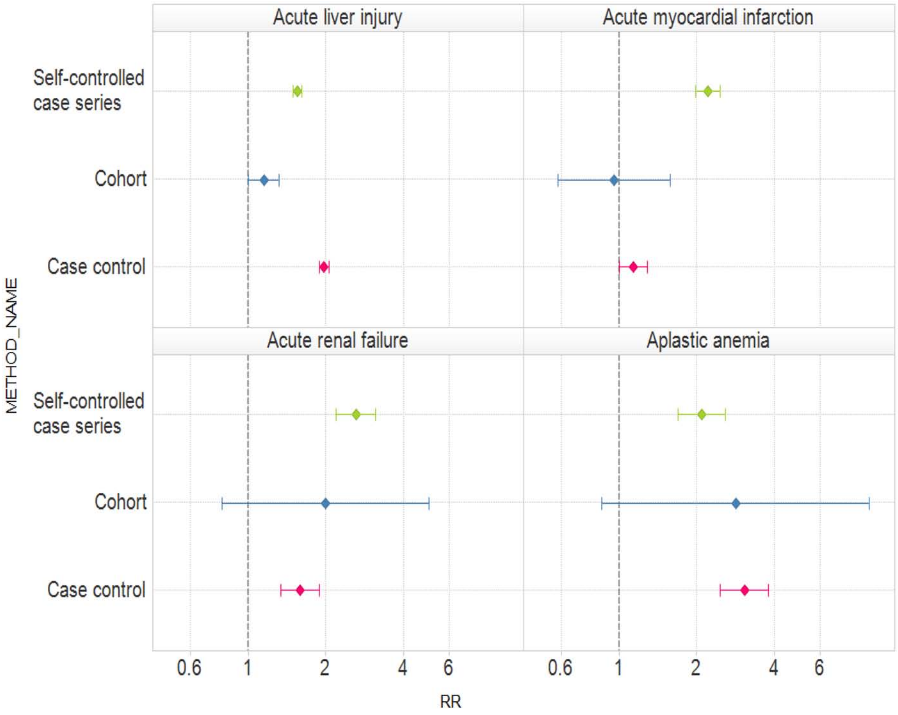

Figure 1. Estimated relative risks (RRs) and 95% confidence intervals for three study designs (self-controlled case series, cohort, and case control) and oral bisphosphonates and four outcomes (acute liver injury, acute myocardial infarction, acute renal failure, and aplastic anemia).

Second, for one particular outcome that is commonly studied in the drug safety literature (acute liver injury), we show the effect of varying one analytic choice at a time for each of the choices of each of the three study designs. Third we show the universe of relative risk estimates provided by all possible analytic choices within each of the three designs for all four outcomes. Finally we show the sensitivity to each analytic choice within each design across all 10 databases included in our study.

Results

Study design choice on four outcomes for bisphosphonates

Figure 1 presents relative risk estimates and 95% confidence intervals for the typical analyses within each of the three study designs and for each of the four outcomes. Table 1 shows the corresponding estimates and corresponding 95% confidence intervals. For acute liver injury, all three study designs produce an estimated relative risk greater than 1 but two are statistically significant (SCCS and case control) while the third is just borderline significant (cohort). For acute myocardial infarction, two of the three relative estimates exceed 1 but one is below 1. SCCSs provides a highly statistically significant positive relative risk, the cohort design provides an estimate indistinguishable from 1 with a relatively wide confidence interval, while the case control design gives a borderline statistically significant positive relative risk. For both acute renal failure and aplastic anemia, all three designs produce relative risk estimates that exceed 1. Two are statistically significant (SCCS and case control) while the third (cohort) is not. Recall that all four outcomes are believed not to be associated with oral bisphosphonates and approximately 95% of all the intervals should include 1. In fact, just 5 of the 12 (42%) intervals include 1. For none of the four outcomes were all three study designs statistically significant in the same direction.

Individual analytic choices within different designs for one outcome

Published studies frequently publish sensitivity analyses that consider particular analytic choices one at a time, comparing the resulting inferences with a baseline analysis. Figure 2 shows the effect of varying individual within-design analytic choices for a case control analysis of oral

Table 1. Estimated relative risks and 95% confidence intervals for three study designs (self-controlled case series, cohort, and case control) and oral bisphosphonates and four outcomes (acute liver injury, acute myocardial infarction, acute renal failure, and aplastic anemia).

Study design	Outcome	Estimated relative risk	95% Confidence interval
Self-controlled	Acute liver injury	1.55	(1.49–1.61)
Cohort	Acute liver injury	1.15	(1.00–1.32)
Case control	Acute liver injury	1.97	(1.90–2.05)
Self-controlled	Acute myocardial infarction	2.22	(1.99–2.47)
Cohort	Acute myocardial infarction	0.96	(0.58–1.58)
Case control	Acute myocardial infarction	1.14	(1.00–1.29)
Self-controlled	Acute renal failure	2.62	(2.20–3.13)
Cohort	Acute renal failure	1.99	(0.79–5.02)
Case control	Acute renal failure	1.59	(1.34–1.90)
Self-controlled	Aplastic anemia	2.09	(1.70–2.58)
Cohort	Aplastic anemia	2.83	(0.86–9.34)
Case control	Aplastic anemia	3.07	(2.47–3.80)

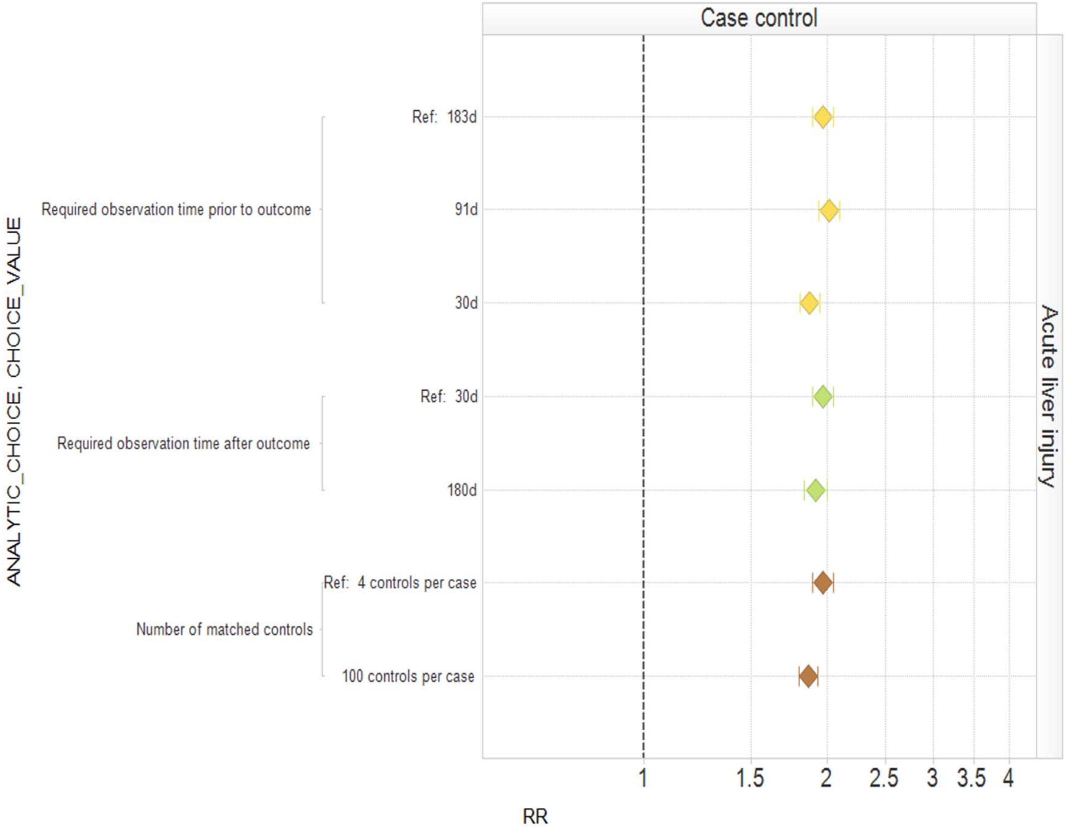

Figure 2. Estimated relative risks (RRs) and 95% confidence intervals for the case control study design with different analytic design choices.

bisphosphonates and acute liver injury around our typical analysis. Varying 'required observation time prior to outcome', 'required observation time after outcome', and 'number of matched controls' individually has little impact on the estimated relative risk and confidence interval. However, since we believe that oral bisphosphonates are not causally related to acute liver injury, all settings

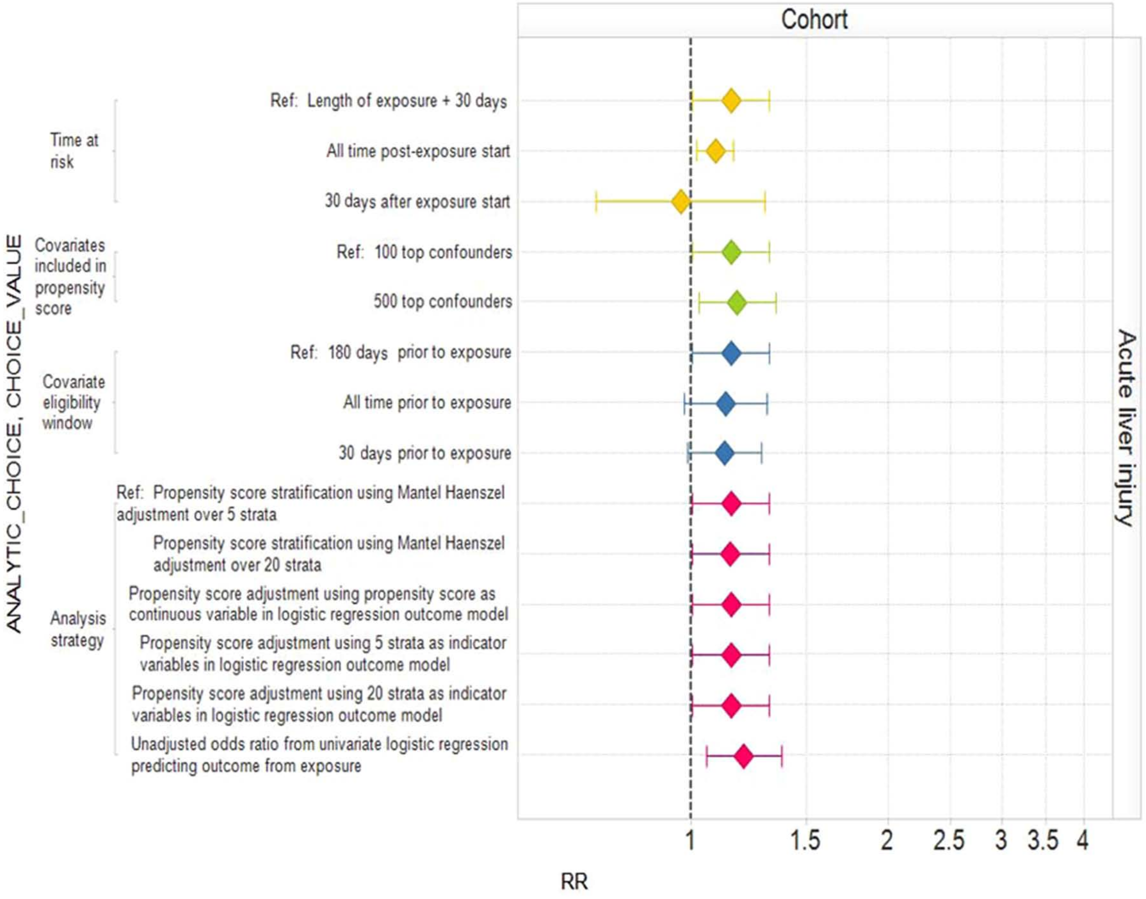

Figure 3. Estimated relative risks (RRs) and 95% confidence intervals for the cohort study design with different analytic design choices.

yield the wrong answer in the sense that the corresponding confidence intervals do not include 1. Figure 3 shows an equivalent analysis for the cohort design varying 'time at risk', 'number of covariates included in propensity score', the 'covariate eligibility window', and the 'analysis strategy' one at a time. Changing the time at risk from the entire length of exposure plus 30 days to ending time at risk 30 days from exposure start moves the estimated relative risk from above 1 to below 1. Other changes affect whether or not the estimated relative risk is statistically significantly above 1 or not. Figure 4 shows the analysis for SCCS. Many of the individual changes show little variation from the typical analysis. However, choosing 'time at risk' to be the first 30 days of exposure more than doubles the estimated relative risk, while exclusion of the index date in the time at risk produces a statistically significant estimate in the opposite direction to the typical (statistically significant) analysis. Appendix 1, (Supplementary data) provides the estimates and confidence intervals corresponding to Figures 2, 3, and 4.

This example shows that the sensitivity revealed by varying one analysis choice at a time can vary from modest (case control) to substantial (SCCS). However, this approach explores a small fraction of all the possible analytic choices. For example, for SCCS, Figure 4 shows just 12 of the 64 possible analyses (four options for two choices and two options for two choices, thus 4 × 4 × 2 × 2 possible combinations), none of which seem unreasonable *a priori*.

Universe of possible estimates for four outcomes for bisphosphonates

Figure 5 shows the relative risk estimates corresponding to all possible analytic choices within each of the three study designs, color coded according to statistical significance. For all four outcomes, the results can vary from statistically significantly above 1 (green) to statistically significantly below 1 (blue), depending on the study design and specific analytic choices. Just within the SCCS design, for all four outcomes, the results can vary from statistically significantly

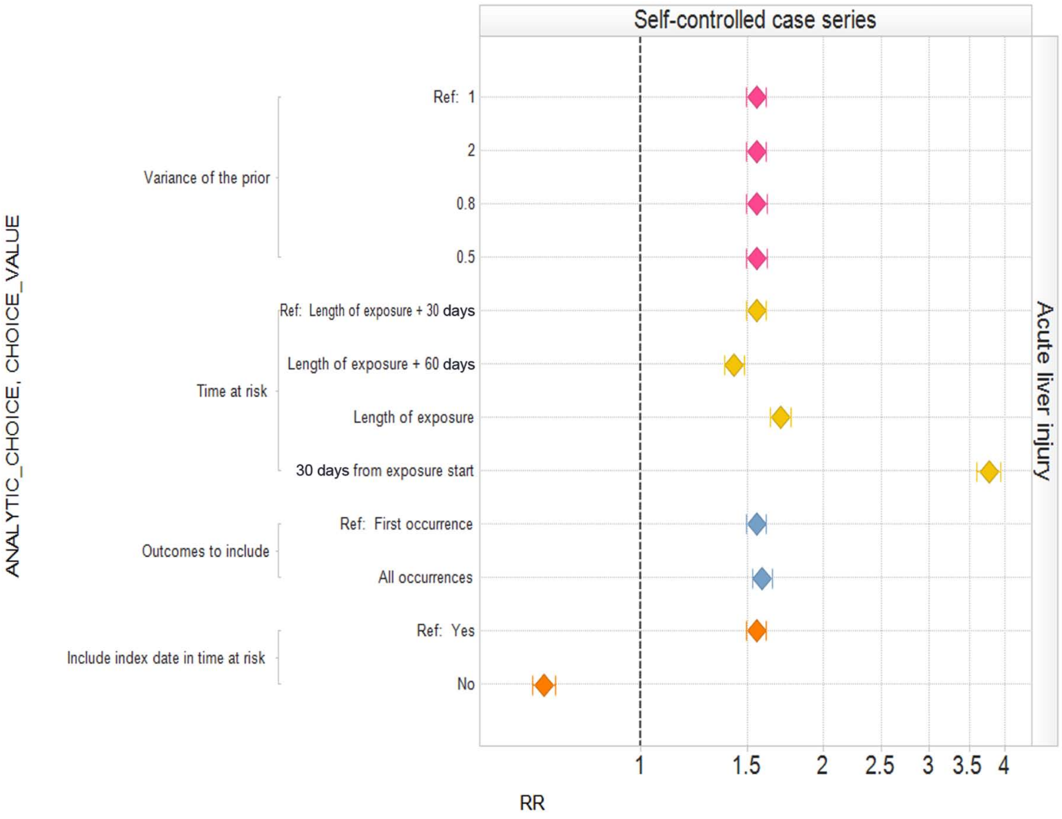

Figure 4. Estimated relative risks (RRs) and 95% confidence intervals for the self-controlled case series study design with different analytic design choices.

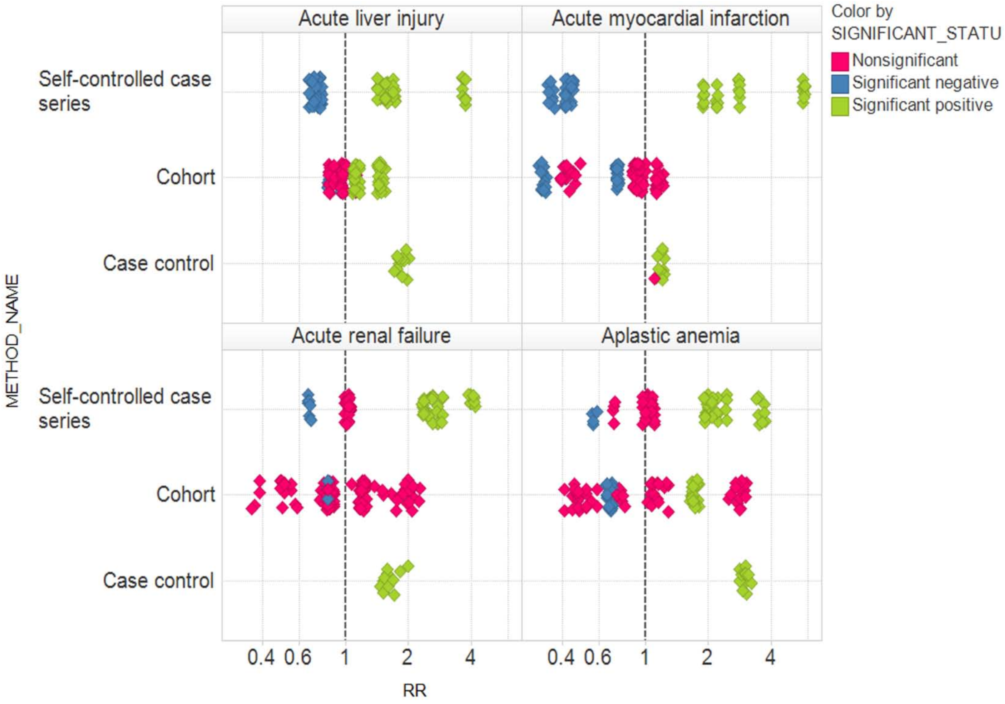

Figure 5. Relative risk (RR) estimates corresponding to all possible analytic choices within each of the three study designs, color coded according to statistical significance.

Table 2. Tenth percentile, 50th percentile (median), and 90th percentile of the distribution of the relative impact of changing a single parameter on the estimated relative risk.

Study design	Parameter	10%ile	50%ile	90%ile
Self-controlled	Time at risk	1.06	1.36	3.20
Self-controlled	Include index date in time at risk	1.03	1.28	3.13
Self-controlled	Outcomes to include	1.01	1.17	2.24
Self-controlled	Variance of the prior	1.00	1.00	1.02
Cohort	Time at risk	1.04	1.25	2.24
Cohort	Covariate eligibility window	1.01	1.05	1.30
Cohort	Analysis strategy	1.00	1.05	1.51
Cohort	Covariates included in propensity score	1.00	1.01	1.08
Case control	Number of matched controls	1.02	1.14	1.63
Case control	Required observation time prior to outcome	1.01	1.05	1.18
Case control	Required observation time after outcome	1.01	1.04	1.19

above 1 to statistically significantly below 1, depending on the specific analytic choices. Within the cohort design, for all four outcomes, relative risk estimates range from above 1 to below 1, with statistical significance in opposite directions for acute liver injury and aplastic anemia. Only for the case control design is some degree of consistency achieved, with all estimates for three of the outcomes being statistically significant in the same direction (above 1). However, we note that all four outcomes here are believed not to be causally related to oral bisphosphonates, and are expected to yield statistically significant associations only 5% of the time if the estimates are truly unbiased.

Heterogeneity due to individual analytic choices

Table 2 illustrates the extent to which analyses are sensitive to individual analytic choices across all 10 databases and across the 53 drug–outcome pairs considered in the OMOP experiment [Stang *et al.* 2010]. For every pair of possible analytic choices, we consider the distribution of the ratio of the two resulting relative risk estimates (larger estimate divided by smaller estimate) for the same drug in the same database. For example, suppose one analytic choice offers three options (A, B, and C). Suppose for a particular drug–outcome pair the resulting relative risk estimates are 1 (A), 2 (B), and 3 (C). Then the three possible ratios are 2/1, 3/1, and 3/2. We compute such ratios across all possible settings for the remaining analytic choices. Table 2 shows the 10th percentile, the 50th percentile, and the 90th percentile of the resulting distributions. For example, for the SCCS method, for 50% of the analyses, changing the setting of 'time at risk' alters the estimated relative risk by a factor of 1.36. For 10% of the analyses, changing the setting of 'time at risk' alters the estimated relative risk by a factor of more than 3.

Table 2 shows that the precise specification of certain analytic choices may not be that critical (e.g. prior variance in the SCCS) but different choices for other parameters (e.g. the time at risk in both the SCCS and cohort designs) may alter the estimated relative risks considerably.

Conclusion

When conducting an observational study, researchers have myriad options for how to design a study. Researchers may develop their own preferences for which designs to use in which circumstances and also may consider some designs less applicable. However, there appears to be little consensus around these preferences within the epidemiology community. For instance, some might argue SCCS should only be used for acute events and transient exposures [Whitaker, 2008], which would not be applicable to oral bisphosphonates; however, others have chosen such a design for this type of exposure [Grosso *et al.* 2009]. Recent efforts have tried to enumerate considerations that go into choosing a study design [Gagne *et al.* 2012], although these stop short of considering analytic choices within study design.

One of the problems in choosing the appropriate design options is that these choices require intimate knowledge of the data, how they are generated, and especially the causal relationship under investigation. For example, the within-design analytic choice of whether to include the index date in the risk window depends on whether we expect bisphosphonates to be capable of causing acute liver injury within a day. But because we do not know this with certainty, we cannot confidently make a design choice that we now know leads to observing either a strong protective or a strong harmful effect. In general, we will be studying a drug precisely because we lack knowledge about its relationship with the outcome of interest, making these design choices dependent on *a priori* assumptions that may not hold true, and will differ between researchers.

Our results show that for one particular type of drug (oral bisphosphonates) and four outcomes, different study designs do lead to qualitatively different conclusions. We chose outcomes that we believe are causally unrelated to oral bisphosphonates and thus most analyses should yield confidence intervals that include 1. Our motivating example analyzed the widely used GPRD database. The results presented in this paper utilize a different outpatient electronic health record database (GE) with its own set of limitations. We have conducted similar analyses for nine other drugs, several different outcomes both related to the drug in question and unrelated, and using nine other large-scale observational databases. The corresponding data are available at http://omop.fnih.org/OMOP2011Symposium. Our findings in every case are qualitatively similar and therefore we do not believe our findings are unique to the specific database, drug, and outcomes reported in this paper. These findings also demonstrate that analyses can be sensitive to the selection of particular combinations of analytic choices. For example, for the four outcomes considered in this paper and across the three study designs, the most positive and most negative relative risk estimates differ by more than one parameter choice in every case, except the self-controlled analysis for acute liver injury.

Sensitivity within study design can also be substantial. Our oral bisphosphonate example illustrates the point that exploring all possible analytic choices provides a more complete sensitivity analysis than varying one particular choice at a time, holding all others constant.

We have focused on just two types of sensitivity analysis, namely sensitivity to choice of study design and sensitivity to analytic choices within study design. Observational studies face many other challenges such as selection bias, unmeasured confounding, and measurement error. While methods for conducting sensitivity analyses with respect to these concerns do exist [Steenland and Greenland, 2004; Greenland, 2005], unfortunately published epidemiologic research has largely ignored these developments [Lash *et al.* 2009]. Methods that appropriately account for uncertainty due to analytic design choice need to become part of standard practice.

Does design matter? These results suggest design matters considerably and that there is a wide range of effect estimates, simply based on the analytic design choices.

Acknowledgement

This article expresses the views of the authors and does not necessarily represent those of their affiliated organizations.

Funding

The Observational Medical Outcomes Partnership is funded by the Foundation for the National Institutes of Health through generous contributions from the following: Abbott, Amgen Inc., AstraZeneca, Bayer Healthcare Pharmaceuticals Inc., Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Johnson & Johnson, Lundbeck Inc., Merck & Co. Inc., Novartis Pharmaceuticals Corporation, Pfizer Inc., Pharmaceutical Research Manufacturers of America (PhRMA), Roche, Sanofi, Schering-Plough Corporation, and Takeda. No funding bodies had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest statement

Drs Ryan, Stang, and Berlin are employees of Johnson & Johnson, and Dr Ryan is a past employee of GlaxoSmithKline.

References

Brixner, D., Said, Q., Kirkness, C., Oberg, B., Ben-Joseph, R. and Oderda, G. (2007) Assessment of cardiometabolic risk factors in a national primary care electronic health record database. *Value Health* 10: S29–S36.

Cardwell, C., Abnet, C., Cantwell, M. and Murray, L. (2010) Exposure to oral bisphosphonates and risk of esophageal cancer. *JAMA* 304: 657–663.

Crawford, A., Cote, C., Couto, J., Daskiran, M., Gunnarsson, C., Haas, K. *et al.* (2010) Comparison of GE centricity electronic medical record database and national ambulatory medical care survey findings on the prevalence of major conditions in the United States. *Popul Health Manag* 13: 139–150.

De Vries, F., De Vries, C., Cooper, C., Leufkens, B. and Van Staa, T. (2006) Reanalysis of two studies with contrasting results on the association between statin use and fracture risk: the general practice research database. *Int J Epidemiol* 35: 1301–1308.

Dixon, W. and Solomon, D. (2011) Bisphosphonates and esophageal cancer – a pathway through the confusion. *Nat Rev Rheumatol* 7: 369–372.

Gagne, J., Fireman, B., Ryan, P., Maclure, M., Gerhard, T., Toh, S. *et al.* (2012) Design considerations in an active medical product safety monitoring system. *Pharmacoepidemiol Drug Saf* 21(Suppl. 1): 32–40.

Gill, J. and Chen, Y. (2008) Quality of lipid management in outpatient care: a national study using electronic health records. *Am J Med Qual* 23: 375–381.

Green, J., Czanner, G., Reeves, G., Watson, J., Wise, L. and Beral, V. (2010) Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case–control analysis within a UK primary care cohort. *BMJ* 341: c4444.

Greenland, S. (2005) Multiple-bias modelling for analysis of observational data. *J R Stat Soc A* 168: 267–306.

Grosso, A., Douglas, I., Hingorani, A., Macallister, R. and Smeeth, L. (2009) Oral bisphosphonates and risk of atrial fibrillation and flutter in women: a self-controlled case-series safety analysis. *PLoS One* 4: e4720.

Hernandez-Diaz, S., Varas-Lorenzo, C. and Garcia Rodriguez, L. (2006) Non-steroidal antiinflammatory drugs and the risk of acute myocardial infarction. *Basic Clin Pharmacol Toxicol* 98: 266–274.

Ioannidis, J. (2005) Why most published research findings are false. *PLoS Med* 2: e124.

Lash, T., Fox, M. and Fink, A. (2009) *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York: Springer.

Loke, Y., Kwok, C. and Singh, S. (2011) Comparative cardiovascular effects of thiazolidinediones: systematic review and meta-analysis of observational studies. *BMJ* 342: d1309.

Mayes, L., Horwitz, R. and Feinstein, A. (1988) A collection of 56 topics with contradictory results in case–control research. *Int J Epidemiol* 17: 680–685.

Meier, C., Schlienger, R., Kraenzlin, M., Schlegel, B. and Jick, H. (2000) HMG-CoA reductase inhibitors and the risk of fractures. *JAMA* 283: 3205–3210.

Overhage, J., Ryan, P., Reich, C., Hartzema, A. and Stang, P. (2011) Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 19: 54–60.

Perrio, M., Waller, P. and Shakir, S. (2007) An analysis of the exclusion criteria used in observational pharmacoepidemiological studies. *Pharmacoepidemiol Drug Saf* 16: 329–336.

Reich, C., Ryan, P., Stang, P. and Rocca, M. (2012) Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *J Biomed Inform* 45: 689–696.

Ryan, P., Madigan, D., Stang, P., Overhage, J., Racoosin, J. and Hartzema, A. (2012) Empirical assessment of analytic methods for risk identification in observational healthcare data: results from the experiments of the observational medical outcomes partnership. *Stat Med* 31: 4401–4415.

Schneeweiss, S., Rassen, J., Glynn, R., Avorn, J., Mogun, H. and Brookhart, M. (2009) High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 20: 512–522.

Smith, G. and Ebrahim, S. (2001) Epidemiology – is it time to call it a day? *Int J Epidemiol* 30: 1–11.

Stang, P., Ryan, P., Dusetzina, S., Hartzema, A., Reich, C., Overhage, J. *et al.* (2011) Health outcomes of interest in observational data: issues in identifying definitions in the literature. *Health Outcomes Res Med* 3: e37–e44.

Stang, P., Ryan, P., Racoosin, J., Overhage, J., Hartzema, A., Reich, C. *et al.* (2010) Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Ann Intern Med* 153: 600–606.

Steenland, K. and Greenland, S. (2004) Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *Am J Epidemiol* 160: 384–392.

Trifiro, G., Pariente, A., Coloma, P., Kors, J., Polimeni, G., Miremont-Salame, G. *et al.* (2009) Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol Drug Saf* 18: 1176–1184.

Van Staa, T., Wegman, S., de Vries, F., Leufkens, B. and Cooper, C. (2001) Use of statins and risk of fractures. *JAMA* 285: 1850–1855.

Von Elm, E., Altman, D., Egger, M., Pocock, S., Gotzsche, P., Vandenbroucke, J. *et al.* (2008) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol* 61: 344–349.

Whitaker, H. (2008) The self controlled case series method. *BMJ* 337: a1069.