Published in final edited form as:

J Cardiothorac Vasc Anesth. 2014 April ; 28(2): 247–254. doi:10.1053/j.jvca.2013.03.014.

The Problem of Controlling for Imperfectly Measured Confounders on Dissimilar Populations: A Database Simulation Study

Robert B Schonberger, MD, MA, Todd Gilbertsen, BS, and Feng Dai, PhD

Department of Anesthesiology, Yale University School of Medicine, New Haven, CT

Abstract

Objective(s)—Observational database research frequently relies on imperfect administrative markers to determine comorbid status, and it is difficult to infer to what extent the associated misclassification impacts validity in multivariable analyses. The effect that imperfect markers of disease will have on validity in situations where researchers attempt to match populations that have strong baseline health differences is underemphasized as a limitation in some otherwise high-quality observational studies. The present simulations were designed as a quantitative demonstration of the importance of this common and underappreciated issue.

Design—Two groups of Monte Carlo simulations were performed. The first demonstrated the degree to which controlling for a series of imperfect markers of disease between different populations taking 2 hypothetically harmless drugs would lead to spurious associations between drug assignment and mortality. The second Monte Carlo simulation applied this principle to a recent study in the field of anesthesiology that purported to show increased perioperative mortality in patients taking metoprolol versus atenolol.

Setting/Participants/Interventions—None.

Measurements and Main Results—Simulation 1: High type 1 error (ie, false positive findings of an independent association between drug assignment and mortality) was observed as sensitivity and specificity declined and as systematic differences in disease prevalence increased. Simulation 2: Propensity score matching across several imperfect markers was unlikely to eliminate important baseline health disparities in the referenced study.

Conclusions—In situations where large baseline health disparities exist between populations, matching on imperfect markers of disease may result in strong bias away from the null hypothesis.

Keywords

misclassification bias; Monte Carlo simulation; database research

In retrospective observational database research, the sensitivity and specificity of comorbidity markers such as ICD-9 codes vary widely^{1,2} and also change as a function of time.³ It is difficult to infer how these inaccuracies impact the validity of such studies, particularly multivariable analyses. We performed a series of Monte Carlo simulations of 2 hypothetical drugs and their effect on mortality to illustrate the degree to which controlling for a series of imperfect markers of comorbidities between 2 populations may lead to

spurious results, specifically in cases in which the populations that are compared have a systematically different prevalence of the underlying confounding comorbidities (ie, when 1 of the populations carries a higher disease burden across several important confounding comorbidities). We then conducted a second Monte Carlo simulation applying this principle to a recent study in anesthesiology purporting to show increased perioperative mortality in patients taking metoprolol as compared to atenolol.⁴

Although the statistical literature has dealt extensively with many of the challenges presented by various types of misclassification,^{5–8} the extent of the impact that imperfect markers of comorbidities will have on validity in situations where researchers attempt to match populations that have strong baseline health differences is poorly understood by most clinicians and is underemphasized as a limitation in some otherwise sophisticated and high-quality observational studies. Moreover, many researchers hold to the rather poorly grounded assumption that administrative codes for major comorbidities such as coronary artery disease, myocardial infarction, and congestive heart failure are “good enough” to allow unbiased comparisons of populations with vastly different prevalences of these comorbidities. The present simulations were designed as a quantitative demonstration of the significant problem that misclassification bias can pose—and in one recent case, has posed—in retrospective observational research using large-scale clinical and administrative databases.

METHODS

Simulation 1

In our first set of simulations, we assigned 2 drugs, A and B, to 10,000 patients each in a 1:1 fashion. The drugs carried no independent effect on mortality. Six binary confounding comorbidities were randomly assigned to each patient with a predefined prevalence of 20% in the drug A group. Prevalence in the drug B group varied across each simulation at 30%, 40%, and 50%. Baseline risk of death was set at 1% for people with no confounding comorbidity. Each additional comorbidity conferred an independent, additive absolute risk of death of 0.5%. Thus, the risk of death for a patient having all 6 comorbidities was 4%.

A series of imperfect ICD-style markers, whose sensitivity and specificity varied from 70% to 100% by 5% intervals, was then created for each of the 6 comorbidities. For each scenario described, 1,000 random data sets were simulated. On each data set, we conducted a propensity score-matched analysis of mortality between drugs A and B using the imperfect markers of the comorbidities as the independent variables in the propensity score models.^{9,10}

For creation of the propensity score-matched cohorts, propensity scores estimating the odds of drug assignment between drug A and drug B were calculated for each simulated patient using a non-parsimonious logistic regression model in which the independent variables were the imperfect markers of comorbidities and the dependent variable was drug assignment. After the creation of these propensity scores, patients were matched as has been described previously in a 1:1 fashion based on the similarity of propensity scores. Matching proceeded via a 5 to 1 greedy matching algorithm in which patients prescribed drug A were matched to patients prescribed drug B first based on 5 digits of the propensity score. For patients not able to match on all 5 digits, an attempt was made to match on 4 digits of the propensity score. This algorithm continued down to 1-digit matching of the propensity score with the remainder of patients who remained unmatched excluded from the propensity score-matched analyses.^{9,11} McNemar's test was then performed on each matched cohort to assess whether Drug A versus Drug B appeared to be independently associated with increased odds of death. For each scenario, the proportion of 1,000 *p* values demonstrating an independent effect of drug assignment on death at the *p* < 0.05 level was calculated and reported as the type I error rate for the several simulated scenarios. Given this arbitrary threshold, a type I

error rate of 0.05 would therefore be expected for a hypothetically perfect database in our simulations.

Historically, large-scale administrative data have suffered more from undercoding (ie, poor sensitivity) than overcoding (ie, poor specificity) of comorbidities.¹² However, more recent work in the popular press has focused on the phenomenon of increasingly widespread overcoding, perhaps as a consequence of financial incentives, that may be associated with the implementation of electronic health records.¹³ We therefore performed a further set of analyses to examine the effects of changes in sensitivity and specificity in isolation, changing only sensitivity or specificity at 5% intervals while holding the other constant at 100% in the effort to demonstrate their relative importance under the simulated database conditions.

Simulation 2

In our second set of simulations, we chose to recreate the propensity score matching protocol used in a recently published study by Wallace et al within the journal, *Anesthesiology*, that compared postoperative mortality among patients taking metoprolol versus atenolol.⁴ In that study, outpatient metoprolol was associated with an increased hazard of postoperative death as compared to atenolol. The outpatient metoprolol population also carried a higher baseline prevalence of administrative markers for peripheral vascular disease (PVD) (45% v 28%), coronary artery disease (CAD) (71% v 48%), prior myocardial infarction (MI) (14% v 4%), and congestive heart failure (CHF) (26% v 8%). As the authors of that study correctly mentioned, matching on imperfect markers of disease may have led to bias, but readers were then reassured that the relevant comorbidity markers likely had sensitivities >80% and specificities >91%.⁴ These evidence-based estimates of comorbidity coding accuracy had previously been found in the database used by Wallace et al and perhaps gave the impression that sensitivities above 80% and specificities above 91% would be sufficiently good to allow for adequate comparison of the very different atenolol and metoprolol populations.

For our second set of simulations, we therefore sought to model whether such sensitivities and specificities would in actuality be “good enough” to successfully match such dramatically different populations. Accordingly, we used published estimates of sensitivity and specificity for the 8 comorbidity markers used in creation of the original propensity-score in the referenced study,^{14–16} with a lower limit set at 80% sensitivity and 91% specificity for any marker with possibly lower published values (actual values used are listed in Table 1). The two remaining variables used in the original propensity score logistic regression were based on age (age itself and “age greater than 60”). In our simulation, age and “age greater than 60” were modeled based on the authors’ reported means and standard deviations⁴ assuming normal distributions. True prevalence of comorbidities was calculated from the reported prevalence of administrative disease markers and published sensitivities and specificities according to the formula: True Prevalence = (Apparent Prevalence + Specificity – 1) / (Sensitivity + Specificity – 1).

For our simulation, comorbidities were randomly assigned to patients based on estimated true prevalence for each population, and corresponding “ICD Comorbidities” were then randomly determined based on the estimated sensitivities and specificities listed in Table 1. For the Monte Carlo simulation, 1,000 iterations of propensity score-matched cohorts were constructed using the same procedures described in simulation 1 above. The outcome measured within each matched cohort was (1) the absolute value of the standardized difference for each of the simulated administrative markers of confounders and (2) the corresponding standardized difference of the true disease confounder variables for the same matched sets. Standardized difference was chosen because it was used by Wallace et al for

assessing covariate balance within their propensity score-matched samples. The referenced study relied on the commonly used value of standardized difference less than 10 to represent adequate balance of confounders in which standardized difference was calculated as 100 times the absolute value of the difference in sample means divided by the pooled standard deviation. Accordingly, “it represent[ed] the difference in means between the two groups in units of standard deviation.”^{17,18}

Sensitivity Analyses

After conducting the above 2 sets of simulations, the authors elected to perform 3 distinct sensitivity analyses to assess the robustness of our primary analysis method of adjusting for confounding by propensity score matching. First, inverse propensity score weighting was used to adjust for confounding¹⁹ instead of matching using propensity scores. In a second sensitivity analysis, a standard multiple logistic regression was used to adjust for 6 comorbidities instead of using propensity score matching. In a third sensitivity analysis, the authors investigated the influence of varying correlation between the comorbidities on the type I error when the authors adjusted for confounding using the propensity score matching method. Instead of 6 independent comorbidities, the authors simulated 3 pairs of correlated comorbidities with different correlation coefficients (−0.8, −0.5, 0, 0.5, 0.8). Baseline risk of death and additional risk conferred from each comorbidity were kept the same as in the first set of simulations.

All simulations in the present manuscript were conducted using SAS software version 9.2 (Cary, NC).

RESULTS

Simulation 1

High type I error (ie, false positive findings of an independent association between drug assignment and mortality) was observed as sensitivity and specificity declined and as systematic differences in disease prevalence increased (Fig 1).

Simulations in which sensitivity and specificity were changed in isolation (Figs 2 and 3) illustrate the previously described principle that in scenarios comparing populations with low prevalence of confounders, specificity of disease markers is a more important determinate of bias than sensitivity.²⁰ For simulation 1, in total the authors performed 63,000 propensity score-matched cohorts representing a population of 1.26 billion simulated subjects.

Simulation 2

Propensity score matching across several imperfect markers would have been unlikely to eliminate important baseline health disparities between groups in contrast to what was claimed in the referenced study.⁴ In our simulated populations, the absolute value of the standardized difference across the 1,000 matched cohorts was consistently elevated, based on the author-defined threshold of standardized difference <10, for several important comorbidities that are known to be powerful confounders of mortality including peripheral vascular disease, prior myocardial infarction, and congestive heart failure (Fig 4, Panel B) despite the appearance of excellent matching on the imperfect administrative markers for those same comorbidities (Fig 4, Panel A). Indeed, all 1,000 simulated cohorts maintained the author-defined threshold standardized difference less than 10 for the full set of imperfect markers of comorbidities while zero of the same 1,000 iterations fulfilled this criterion on simulated true comorbidity status.

Sensitivity Analyses

The results of our sensitivity analyses demonstrated similar findings for changes in type I error rates over the different choices of sensitivity and specificity of each comorbidity regardless of propensity score controlling method (Fig 5). We also found that propensity score-based methods were not noticeably superior or inferior to a straightforward logistic regression model in which all 6 comorbidities were included as covariates, which was likely due to the large event per variable (EPV) ratios in our simulations (more than a recommended EPV ratio of 10 for logistic regression).²¹ The interested reader is referred to the relevant references^{22,23} for details on this issue.

When comorbidities were correlated, similar trends for changes of type I error were identified (Fig 6). At each scenario of simulated sensitivity and specificity values, type I error rates, when there were positive correlations between 2 comorbidities, were lower than those when there were no correlations and further lower than those when there were negative correlations between 2 comorbidities. This suggests that when 2 comorbidities' effects on outcome are in the same direction and positively correlated with each other, the propensity score method is more efficient in adjusting for the comorbidities than when there is negative correlation. Given these results, some considerations for interaction effects of different comorbidities in deriving propensity scores when using real clinical data may be advisable.

DISCUSSION

The results of Simulation 1 demonstrate that the distortions that stem from misclassification of comorbidity status grow as the populations one wishes to compare become more dissimilar. Such distortions are a manifestation of the well-known statistical phenomenon that the positive and negative predictive values of an imperfect comorbidity marker will change based on the underlying prevalence of the true comorbidity in the population. For example, the positive predictive value of an imperfect code for congestive heart failure will be much higher in a population with widespread heart failure than in another population with a low prevalence of heart failure. While the relationships among predictive values, disease prevalence, and statistical accuracy are well known to researchers in a general sense, their distortional effect in multivariable analyses comparing databases of sicker and healthier populations remains underappreciated. Such distortion creates an inherent bias toward mortality in the baseline sicker population that may strongly prejudice results away from the null hypothesis. Our first set of simulations is meant as a graphical reminder to the reader that the requirement of highly accurate comorbidity markers becomes increasingly important as the baseline health differences one is trying to control for become more dramatic.

In Figures 2–3 of Simulation 1, we found that given the assumed prevalence of our comorbidities, specificity was more important to type I error than sensitivity. Some statistical sources still make the erroneous statement that in situations of nondifferential misclassification as in the simulated scenarios, the associated bias will necessarily be toward the null hypothesis,²⁰ but the truth is that the direction of bias will depend on the pattern of prevalence differences between populations and the relationship between the several confounders and outcome.^{24,25}

In our several scenarios in Simulation 1, we compared two populations that demonstrated dramatically different disease prevalence across several powerful confounders of the outcome of interest, and some readers may feel that such a scenario was overly contrived, perhaps lacking clear applicability to real-world data. Accordingly, Simulation 2 was intended as a practical application of the principles outlined in Simulation 1. As described in the results to Simulation 2, we found that given the estimated sensitivity and specificity of

disease markers in the referenced study, Wallace et al's study populations were simply too different to match to the rigorous standards that they themselves suggested (ie, standardized difference <10) given the database and matching methods they were using. The authors hasten to add that while they clearly put into question the validity of one of the matching protocols and associated conclusions in Wallace et al's manuscript, they do not claim to definitively contravene their findings. First, the authors focused on only 1 portion of the sophisticated and provocative research performed by Wallace et al.⁴ In addition, there was no access to the actual propensity score logistic regression model used by the Wallace et al authors or to their multivariable Cox regression models. The authors were therefore unable to make a reliable estimate about the effects that the persistent covariate imbalance modeled herein would have had on Wallace et al's final conclusions. Moreover, all of the above simulations assumed independent, nondifferential misclassification of disease markers,²⁵ which are acknowledged to be a simplification of the true status of any actual series of administrative markers across different populations. Nevertheless, given the large baseline health disparities between the populations in the referenced article taking metoprolol and atenolol, specifically regarding the prevalence of PVD, CAD, MI, and CHF, it would have been extraordinarily difficult to match the 2 populations adequately on these significant comorbidities using a propensity score regression model created from age-based variables and administrative disease markers alone.

We do not mean to imply that misclassification bias, sometimes also referred to as information bias, is the only type of bias that afflicts papers attempting to compare group assignment on outcome. Particularly germane to large observational studies, other sources of bias include selection bias,²⁶ confounding by indication,²⁷ and time-dependent confounding by treatment,²⁸ each of which have played an important role in other work that has attempted to define the treatment effects of particular drugs on an outcome of interest. Moreover, as the sensitivity analyses above demonstrate, propensity score matching is not uniquely susceptible to misclassification bias. Indeed, with sufficient event per variable ratios, it is neither better nor worse than other uses of propensity scores or indeed a simple multiple logistic regression model in failing to compensate for misclassification bias. We simply chose propensity score matching in Simulation 1 in order to maintain analytical continuity with the methods used in Simulation 2, which themselves were simply meant to emulate a relevant study in the recent anesthesiology literature. On the contrary, misclassification bias is a factor that depends not on the analytical method used but rather on (1) the foibles of the database used, (2) the comorbidity prevalence differences of the populations one attempts to compare, and (3) the strength and patterns of the relationships among confounding comorbidities and outcome.

It is important to acknowledge several limitations of our study. First, we only considered 6 comorbidities in our simulations, which might be seen as somewhat parsimonious compared to most administrative analyses in which many more covariates are routinely included in creating propensity scores.^{29,30} In addition, the effects of those covariates on the dependent outcome would naturally be more complicated than the co-additive effects we assumed in our simulations. While we chose to model a convenient simplification of the relationships between comorbidities and outcomes, this does not change the fundamental relationships among predictive values, disease prevalence, and statistical accuracy that are the central foundations of our simulations of misclassification bias. Also, we defined the proportion of 1000 iterations in which a *p* value was less than 0.05 as the type I error rate under each simulated scenario. The arbitrary choice of 1,000 iterations for each scenario was a combination of habitual allegiance to a base 10 system, and what we felt to be a reasonable compromise between rigor and computational time. Although we agree that more iterations (eg, 10,000) would yield increasingly accurate estimates of type I error, we argue that 1,000 iterations was sufficient given our empirical results in which we constantly observed

nominal values of type I error of around 0.05 when each comorbidity was assumed to have perfect (ie, 100%) sensitivity and specificity. Finally, we only presented one estimate (from 1,000 iterations) of type I error rate for each set of database assumptions. It would certainly be possible, if not particularly useful, to repeat any particular simulation for multiple times (eg, $100 \times 1,000$ iterations) to calculate the standard deviations of the estimates. However, it should be remembered that our quantitative estimates of type I error would still remain only a reflection of other arbitrary choices within this manuscript regarding the simulated relationships among comorbidities and outcome.

Importantly, none of the above limitations impact the problem whose illustration is the primary purpose of this manuscript – namely, the problem of misclassification bias, which stems from a set of simple and widely understood statistical principles. The core principles at work are that the positive and negative predictive values of markers of confounders are dependent upon the true prevalence of confounders in the population. This fact makes it fundamentally impossible to use many presently available databases to adequately compare populations that are simply too different from each other at baseline, whether those populations are groups of people taking different medications, groups of hospitals treating different patient populations, or some other object of risk-adjusted analysis. Defining how big a difference between groups is “just too different” depends not on the multivariable method used but on the inaccuracies of the database in question and the number, relationship, and effects of the imperfectly coded confounding comorbidities on the particular outcome of interest. Misclassification bias is therefore an ever-present, and occasionally insurmountably present, function of the interaction between a researcher who desires to ask a question and a given database that may simply not be up to the task.

The present study is not intended to discredit all observational studies that use imperfect markers of disease, to criticize the use of administrative codes in general, or to cast aspersions on the extraordinarily rich VA database that has been used both by Dr. Wallace and ourselves in prior anesthesiology research.^{31,32} On the contrary, we first wish to emphasize that all markers will carry some inherent error within them. Secondly, administrative markers in particular carry several advantages over other types of data. These advantages include broad penetrance within large EHR databases, relevance across the breadth of medical specialties, and in many instances, good reliability. Indeed, in some situations, the use of different ICD-9 codes to cross-reference other administrative data has been appropriately advocated as a method for decreasing some forms reporting bias.³³ Other investigators have maintained validity simply by picking their database well. For example, in their study of perioperative risk factors of acute kidney injury,³⁴ Kheterpal et al, used validated data from the American College of Surgeons National Surgical Quality Improvement Project that were high-quality and thus not particularly vulnerable to the dangers outlined in our simulations. Another option to deal with the dangers of comparing different populations using imperfect markers of disease is to include important non-administrative-marker-based variables that are highly correlated with the outcome of interest, as was done in our own recent study of perioperative heart rates among users of atenolol and metoprolol where a baseline primary care heart rate for each subject was included in our multivariable models.¹¹ Rather than undermine observational databases in general, it is our hope that the simulations conducted herein will alert authors to a more critical appraisal of the potential limitations that should be considered in designing and reporting administrative database studies of starkly different population groups.

In conclusion, motivated by the desire to understand the effects that imperfect markers of disease will have on analyses of administrative data, our simulations emphasize the need for caution in conducting, reporting, and interpreting research based on controlling for imperfect markers of powerful confounding conditions when comparing populations with strong

differences in baseline health status. Particularly in such cases of large baseline health differences, researchers should emphasize—or where possible, model the impact of—bias due to imperfectly measured confounders. With such guidance, researchers can improve the quality and accurate interpretation of the growing body of important scientific literature that is based on the analysis of large-scale administrative databases in anesthesiology and throughout medicine.

References

1. Bazarian J, Veazie P, Mookerjee S, et al. Accuracy of mild traumatic brain injury case ascertainment using ICD-9 codes. *Acad Emerg Med*. 2006; 13:31–38. [PubMed: 16365331]
2. Golomb M, Garg B, Williams L. Accuracy of ICD-9 codes for identifying children with cerebral sinovenous thrombosis. *J Child Neurol*. 2007; 22:45–48. [PubMed: 17608305]
3. Bharmal MF, Weiner M, Sands LP, et al. Impact of patient selection criteria on prevalence estimates and prevalence of diagnosed dementia in a Medicaid population. *Alzheimer Dis Assoc Disord*. 2007; 21:92–100. [PubMed: 17545733]
4. Wallace AW, Au S, Cason BA. Perioperative β -blockade: Atenolol is associated with reduced mortality when compared to metoprolol. *Anesthesiology*. 2011; 114:824–836. [PubMed: 21372680]
5. Bross I. Misclassification in 2 X 2 tables. *Biometrics*. 1954; 10:478–486.
6. Copeland KT, Checkoway H, McMichael AJ, et al. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977; 105:488–495. [PubMed: 871121]
7. Diamond EL, Lilienfeld AM. Effects of errors in classification and diagnosis in various types of epidemiological studies. *Am J Public Health*. 1962; 52:1137–1144.
8. Goldberg JD. The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *J Am Stat Assoc*. 1975; 70:561–567.
9. Parsons, L. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. Proceedings of the Twenty-Sixth Annual SAS® Users Group International Conference; 2001; Cary, NC: SAS Institute Inc; 2001. p. 214-226.
10. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
11. Schonberger RB, Burg M, Feinleib J, et al. Atenolol is associated with lower day of surgery heart rate as compared to long and short-acting metoprolol. *J Cardiothorac Vasc Anesth*. 2012; 27:298–304. [PubMed: 22889605]
12. Austin PC, Tu JV, Alter DA, et al. The impact of under coding of cardiac severity and comorbid diseases on the accuracy of hospital report cards. *Med Care*. 2005; 43:801–809. [PubMed: 16034294]
13. Abelson, R.; Creswell, J.; Palmer, G. The New York Times. Sep 21. 2012 Medicare bills rise as records turn electronic.
14. Crothers K, Goulet JL, Rodriguez-Barradas M, et al. Decreased awareness of current smoking among health care providers of HIV-positive compared to HIV-negative veterans. *J Gen Intern Med*. 2007; 22:749–754. [PubMed: 17503106]
15. Justice AC, Lasky E, McGinnis KA, et al. Medical disease and alcohol use among veterans with human immunodeficiency infection: A comparison of disease measurement strategies. *Med Care*. 2006; 44:S52–S60. [PubMed: 16849969]
16. Szeto HC, Coleman RK, Gholani P, et al. Accuracy of computerized outpatient diagnoses in a veterans affairs general medical clinic. *Am J Manag Care*. 2002; 8:37–43. [PubMed: 11814171]
17. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med*. 2008; 27:2037–2049. [PubMed: 18038446]
18. Flury B, Reidwyl H. Standard distance in univariate and multivariate analysis. *Am Stat*. 1986; 40:249–251.
19. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998; 17:2265–2281. [PubMed: 9802183]
20. Szklo, M.; Nieto, FJ. Epidemiology: Beyond the Basics. 2. Canada: Jones and Bartlett; 2007.

21. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996; 49:1373–1379. [PubMed: 8970487]
22. Biondi-Zoccai G, Romagnoli E, Agostoni P, et al. Are propensity scores really superior to standard multivariable analysis? *Contemp Clin Trials*. 2011; 32:731–740. [PubMed: 21616172]
23. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *J Clin Epidemiol*. 2001; 54:979–985. [PubMed: 11576808]
24. Fleiss, JL.; Levin, B.; Paik, MC. Statistical methods for rates and proportions. 3. United States: Wiley-Blackwell; 2004.
25. Rothman, KJ.; Greenland, S. Modern Epidemiology. 2. Philadelphia: Lippincott Williams & Wilkins; 1998.
26. Krumholz HM. Registries and selection bias: The need for accountability. *Circ Cardiovasc Qual Outcomes*. 2009; 2:517–518. [PubMed: 20031886]
27. Han L, Kim N, Brandt C, et al. Antidepressant use and cognitive deficits in older men: Addressing confounding by indications with different methods. *Ann of Epidemiol*. 2011; 22:9–16. [PubMed: 22037381]
28. Gerhard, T. Time-Dependent Confounding in Antihypertensive Drug Studies [dissertation]. Gainesville, FL: University of Florida; 2007. at http://ufdcimages.uflib.ufl.edu/UF/E0/01/86/80/00001/gerhard_t.pdf, 2007 [Accessed 2/5, 2013]
29. Ellenberger C, Tait G, Beattie WS. Chronic beta blockade is associated with a better outcome after elective noncardiac surgery than acute beta blockade: A single-center propensity-matched cohort study. *Anesthesiology*. 2011; 114:817–823. [PubMed: 21427537]
30. Wallace AW, Au S, Cason BA. Perioperative beta-blockade: Atenolol is associated with reduced mortality when compared to metoprolol. *Anesthesiology*. 2011; 114:824–836. [PubMed: 21372680]
31. Schonberger RB, Burg MM, Holt NF, et al. The relationship between day-of-surgery and primary care blood pressure among veterans presenting from home for surgery: Is there evidence for anesthesiologist-initiated blood pressure referral? *Anesth Analg*. 2012; 114:205–214. [PubMed: 22075017]
32. Wallace AW, Au S, Cason BA. Association of the pattern of use of perioperative β -blockade and postoperative mortality. *Anesthesiology*. 2010; 113:794–805. [PubMed: 20864832]
33. Larach M, Brandom B, Allen G, et al. Administrative databases: Are they useful for clinical analyses? *Anesthesiology*. 2009; 111:450. [PubMed: 19625815]
34. Kheterpal S, Tremper KK, Heung M, et al. Development and validation of an acute kidney injury risk index for patients undergoing general surgery: Results from a national data set. *Anesthesiology*. 2009; 110:505–515. [PubMed: 19212261]

Fig 1.
Type I error rates under changing sensitivity, specificity, and prevalence disparity.

Fig 2.

Type I error in the setting of perfect specificity assuming a positive relationship between confounders and the outcome of interest.

Fig 3. Type I error in the setting of perfect sensitivity assuming a positive relationship between confounders and the outcome of interest.

Fig 4. Standardized Difference <10 was the author-defined threshold for adequate covariate balance (Vertical red lines in Panel A and Panel B). (A) All administrative markers of confounding comorbidities in the propensity-score model appeared well-matched. (B) The combination of large prevalence differences and imperfect sensitivity and specificity led to persistent comorbidity imbalance for the true conditions CAD, CHF, PVD, and MI in the same matched groups.

Fig 5. Sensitivity analysis demonstrating that Type I error was approximately invariant to the method of multivariable adjustment used in Simulation 1.

Fig 6.

Sensitivity analysis demonstrating the effects of different correlations among comorbidities on Type I error in Simulation 1. The propensity-score matching method appears more efficient at accommodating positive comorbidity correlations as compared to negative comorbidity correlations. Propensity-score matched analyses may benefit from a consideration of interaction effects among comorbidities included in the creation of the underlying propensity scores.

Table 1

Comorbidity	Sensitivity	Specificity	Observed Prevalence in Atenolol Group	Observed Prevalence in Metoprolol Group
Coronary Artery Disease	0.89	0.98	0.48	0.71
Myocardial Infarction	0.80	0.98	0.04	0.14
Congestive Heart Failure	0.87	1.00	0.08	0.26
Peripheral Vascular Disease	0.80	0.99	0.28	0.45
Diabetes	1.00	0.99	0.30	0.36
Hypercholesterolemia	0.95	0.94	0.54	0.51
Hypertension	1.00	0.94	0.85	0.75
Smoking	0.82	0.93	0.23	0.20

NOTE: Estimated Sensitivity, Specificity and Observed Prevalence of the Eight Comorbid Disease Markers Used for Simulation 2 and in the Referenced Article that Attempted to Compare Perioperative Mortality Between Patients Taking Outpatient Atenolol and Metoprolol.