Prospective cohort studies of newly marketed medications: Using covariate data to inform the design of large-scale studies

Jessica M. Franklin\textsuperscript{a}, Jeremy A. Rassen\textsuperscript{a}, Dorothee B. Bartels\textsuperscript{b,c}, Sebastian Schneeweiss\textsuperscript{a}

\textsuperscript{a.)} Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA  
\textsuperscript{b.)} Department of Global Epidemiology, Boehringer Ingelheim GmbH, Ingelheim, Germany  
\textsuperscript{c.)} Institute for Epidemiology, Social Medicine and Health Systems Research, Hannover Medical School, Hannover, Germany
Series Editors:

Sebastian Schneeweiss, MD, ScD
Jerry Avorn, MD
Robert J. Glynn, ScD, PhD
Niteesh K. Choudhry, MD, PhD
Jeremy A. Rassen, ScD
Josh Gagne, PharmD, ScD

Contact Information:

Division of Pharmacoepidemiology and Pharmacoeconomics
Department of Medicine
Brigham and Women’s Hospital and Harvard Medical School
1620 Tremont St., Suite 3030
Boston, MA 02120
Tel: 617-278-0930 Fax: 617-232-8602
Original Article

Prospective cohort studies of newly marketed medications: Using covariate data to inform the design of large-scale studies

Jessica M Franklin*1, Jeremy A Rassen1, Dorothee B Bartels2,3, and Sebastian Schneeweiss1

1 Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
2 Department Global Epidemiology, Boehringer Ingelheim GmbH, Ingelheim, Germany
3 Institute for Epidemiology, Social Medicine and Health Systems Research, Hannover Medical School, Hannover, Germany

January 24, 2013

Running head: Prospective cohort studies of newly marketed medications

Conflicts of Interest and Source of Funding: This research was funded by an investigator-initiated contract from Boehringer Ingelheim GmbH to the Brigham and Women's Hospital.

Acknowledgements: The authors would like to thank Tim Waterboer, Eva Kleine, Kristina Zint, and Nils Schoof for helpful comments on earlier drafts of this manuscript and SeoYoung Kim for providing expertise in the example study.

* Address correspondence to Dr. Jessica Franklin, 1620 Tremont St., Suite 3030, Boston, MA 02120; email: JMFranklin@partners.org; ph: 617-278-0675; fax: 617-232-8602.
Original Article

Prospective cohort studies of newly marketed medications: Using covariate data to inform the design of large-scale studies

January 24, 2013
Abstract

Background: Non-randomized safety and effectiveness studies are often initiated immediately following the approval of a new medication, but patients prescribed the new medication during this period may be substantially different from those receiving an existing comparator treatment. Restricting the study to comparable patients after data have been collected can be inefficient in prospective studies with primary collection of outcomes.

Methods: We discussed methods for evaluating covariate data in order to assess the comparability of patient populations and identify patient subgroups that are not comparable prior to initiating a large-scale comparative study. We demonstrated methods in an example study of Cox-2 inhibitors during their postmarketing period (1999-2005) versus nonselective nonsteroidal anti-inflammatory (ns-NSAID) drugs.

Results: Covariate mean differences and the $C$-statistic from the propensity score model indicated that covariate imbalance was reduced over time but remained high following the approval of Cox-2 inhibitors. In contrast, estimated bias due to imbalance was not reduced over time. The distribution of Cox-2 and ns-NSAID patients across strata of covariates and the propensity score indicated that overlap was generally strong on all covariates across the study period, resulting in low estimated model extrapolation bias.

Conclusions: A sequential study design that uses pilot covariate data to evaluate treatment selection can improve the efficiency of large-scale outcome studies with primary data collection by assessing only comparable populations. However, the benefit of such an approach is highly dependent on the specific treatment situation. In the example study, substantial patient restriction would not be required if covariate distributions are similar to the pilot data.
Introduction

Non-randomized safety and effectiveness studies are often initiated immediately following the launch of a new medication\textsuperscript{1,2}. These studies are focused on understanding the effects of drugs as used in routine care by populations that are often excluded in pre-approval studies, such as children, the elderly, women of child-bearing age, or patients with many co-morbidities or co-medications\textsuperscript{3}. In order to infer causation from these studies, alternative treatment strategies must be compared in patients that are similar with respect to characteristics that influence their baseline risk of outcome.

In non-randomized studies, equitable comparisons are achieved by measuring and adjusting for factors that influence outcome through one of several available methods, including matching, stratification, or regression on covariates, on the propensity score (PS), or on a disease risk score\textsuperscript{4-9}. Each of these methods can produce unbiased estimates of the average treatment effect (ATE) in the subgroup of patients that are comparable, but they generally require that patients that are not comparable with respect to measured characteristics be explicitly or implicitly removed from the study sample prior to estimation. For example, matching or trimming the tails of an observed PS distribution in cohort studies has been shown to improve subsequent confounding adjustment and effect estimation\textsuperscript{10}.

In the context of newly marketed medications, patients prescribed the new medication may often be sicker, be more likely to have already failed on existing therapy, or have better access to healthcare compared with patients receiving an existing comparator treatment\textsuperscript{11}. Therefore, treatment groups may be largely incomparable, meaning that many patients must be removed from the analysis of treatment effects for an appropriate
comparison. In electronic healthcare data, this restriction will reduce study size, but it will likely not directly impact the cost of completing the study. However, if investigators recruit, contact, interview, or examine patients, then restriction to comparable patients after all data have been collected can be costly and inefficient. In this setting, it may be more efficient to focus study resources on patients that are comparable, either by restricting outcome collection to a subgroup of patients or by waiting to collect outcomes until the new treatment is prescribed less selectively. Since collecting data on covariates is often much less expensive than following patients for outcomes, investigators may wish to evaluate patient characteristics and treatment choice to aid in design choices for a planned large-scale prospective study with primary data collection.

The objective of this article is to discuss methods for the evaluation of covariate data in order to assess the comparability of patient populations and identify patient subgroups that are not comparable prior to initiating a large-scale comparative study. We focused on the case when covariate data is available for a representative sample of the patient population of interest prior to the initiation of primary data collection on outcomes. We demonstrated methods in an example study of Cox-2 inhibitors during their early postmarketing period versus nonselective nonsteroidal anti-inflammatory drugs (ns-NSAIDs).
Methods

Study design

We assume a non-randomized cohort study scenario in which the study goal is to estimate the comparative safety and effectiveness of a newly marketed medication versus an existing alternative treatment for the same clinical indication. Prior to study initiation, investigators should identify all relevant factors that influence the study outcome and that can be measured at baseline. Investigators then identify patients that have initiated either of the two treatments and record covariate information and treatment status. Based on these characteristics, investigators have the following options: 1) initiate the large-scale study, collect data on all newly enrolled patients, and analyze a subset of enrolled patients; 2) initiate the large-scale study but modify study inclusion criteria to increase the comparability of treatment groups; or 3) delay the large-scale study until the new therapy is prescribed less selectively.

The goal of this decision is to efficiently utilize study resources by collecting outcome data primarily on patients that can be used in comparative analyses. The decision is not informed by the large-scale study result, which is unknown at the time. If investigators choose to delay the large-scale study, the decision can be revisited (Figure 1).

In this design, patients that initiate treatment soon after drug approval are enrolled in the pilot phase and assessed for covariates. Interim analyses are conducted using data on newly enrolled patients to determine whether to begin the large-scale study and follow patients for outcomes. This design assumes that the trends in treatment choice observed in pilot phase patients will extend to the future patients available for the large-scale study.
Assessing covariate balance

Covariate balance occurs when the distributions of covariates are identical in both treatment groups. If all prognostic factors for outcome are balanced between treatment groups, then covariate adjustment is not necessary for unbiased estimation of treatment effects, although adjustment is still valid and can improve precision\textsuperscript{13}. If some or all covariates that predict outcome are not balanced, then those covariates should be adjusted for. In non-experimental medication studies, covariates almost always display some imbalance, and investigators generally assume that adjustment for covariates will be necessary\textsuperscript{14-16}. However, assessing covariate balance is a natural first step for evaluating the proposed cohort study as it provides information on the strength of confounding and the relative importance of the available covariates.

Past work has shown that the $C$-statistic from the PS model predicting treatment choice is useful for assessing potential bias due to covariate imbalance\textsuperscript{17}. In addition, examining absolute differences in covariate means between treatment groups can help researchers determine which covariates are primarily responsible for the differences between groups. We can also calculate the estimated bias of treatment effects based on the observed covariate data and assumed associations between covariates and outcome. This estimate provides a more interpretable measure of imbalance as well as a method for incorporating information on the association between each covariate and the outcome.

Let $X_i = (X_{1i}, X_{2i}, ..., X_{Ci})'$ be a vector of covariate values for study patient $i$, and let $T_i$ be the corresponding indicator of treatment assignment ($T_i = 1$ indicates the new treatment, $T_i = 0$ indicates the comparator). For simplicity, we assume that all patients will be followed for the same amount of time and the outcome ($Y_i$) is a Poisson event count.
that is related to treatment and covariates through the model:

\[
\log \{E(Y_i|T_i, X_i)\} = \beta_0 + \beta_T T_i + \beta X_i
\]

where \(E(Y_i|T_i, X_i)\) is the expectation of \(Y_i\) given \(T_i\) and \(X_i\), \(\beta = (\beta_1, \beta_2, \ldots, \beta_C)\) is a vector of coefficients on the covariates in \(X_i\), and \(\beta X_i = \sum_{c=1}^{C} \beta_c X_{ci}\). In this model, \(e^{\beta_T}\) is the true rate ratio (RR) treatment effect. Let \(\hat{r}_T\) denote the crude event rate in treatment group \(T\). As shown in the Web Appendix, the bias of the RR treatment effect estimator can be approximated by:

\[
\text{Bias}\left(\frac{\hat{r}_1}{\hat{r}_0} \bigg| X_i\right) = e^{\beta_T} \left[\frac{m_1}{m_0} - 1\right]
\]

where \(m_t\) is the mean of \(e^{\beta X_i}\) in patients with \(T_i = t\). To remove the dependence on the true treatment effect, we can simply divide the formula above by \(e^{\beta_T}\) to provide an estimator of the percent bias.

Although the calculations above assume uniform follow-up time and constant baseline hazard for all patients, they provide a reasonable approximation when these assumptions are violated in a non-systematic way. In addition, the values of \(\beta\) used to estimate bias can be approximate based on existing knowledge; their primary utility is in distinguishing between strong and weak predictors of each outcome. To evaluate the expected bias after complete adjustment for specific covariates, one can set the RR for those covariates to 1 and recalculate bias. Also, this calculation only accounts for bias due to imbalance on measured covariates; bias due to model misspecification in the presence of treatment effect heterogeneity is considered in the next section.
Assessing covariate overlap

Covariate overlap refers to the range of covariate values observed within each treatment group, and perfect overlap indicates that for every treated patient, there exists at least one comparable untreated patient and vice versa. Overlap may be considered univariately (for example, do older or younger patients never receive the new treatment?) or multivariately (are there any combinations of patient characteristics that never occur in one treatment group?). Overlap is a less strict criterion than balance; good balance on a covariate implies good overlap, but a lack of balance does not necessarily indicate a lack of overlap. Web Appendix Figure 1 shows example age distributions that highlight the difference between balance and overlap.

The implications of covariate overlap depend on the analysis approach that will be used in the large-scale study. If investigators plan to remove patients that lie in regions of nonoverlap from the analysis, then overlap in the pilot data provides information on what proportion of patients will be removed. Investigators can remove nonoverlapping patients by directly trimming patients based on their PS value10 or through stratification on the PS9, where strata are chosen to achieve covariate balance within each stratum. With these approaches, the primary impact of overlap is on the precision of the resulting treatment effect estimator, and investigators can evaluate this impact by examining the proportion of patients that lie in regions of nonoverlap on the PS.

Although trimming is commonly performed in secondary data analyses, this technique may be undesirable or impossible if data collection is expensive or if a regulatory agency requires that all patients enrolled in the large-scale study be used in comparative analyses. If investigators do not plan to remove nonoverlapping patients in the analysis of
the large-scale study and will instead use an analysis approach that includes all patients, such as multivariate regression on treatment and covariates or weighting on the PS, then overlap on covariates is required to ensure an unbiased estimator of treatment effect, since nonoverlap in this case can lead to inappropriate extrapolation.

Investigators can assess the multidimensional overlap on covariates by graphically comparing the density of the estimated PS across treatment groups and by calculating the number of exposed and unexposed patients in strata of the PS. Similar methods can be used to assess the overlap on each individual covariate. We recommend these approaches over strict comparisons of covariate ranges in each treatment group because inappropriate extrapolation can occur even if the ranges are identical. For example, an investigator may observe that each treatment group has a maximum age of 105 years but that only a few patients over 90 years old received the new medication. Thus, the age range in identical across treatment groups, but the small number of very old patients on the new medication may exert undue influence on treatment effect estimates.

With assumptions about the size of the treatment effect and the amount of heterogeneity, we can also estimate the approximate bias due to inappropriate model extrapolation given an observed covariate distribution. Specifically, we assume outcome is a Poisson event related to treatment and a single covariate, $X$, through the model

$$\log\{E(Y_i|T_i, X_i)\} = \beta_0 + \beta_T T_i + \beta_X X_i + \beta_H T_i X_i$$

In this case, the true RR treatment effect is heterogeneous, given by $e^{\beta_T + \beta_H x}$ for a patient with $X = x$. The ATE is the expectation of this quantity over the values of $X$ in the population. If treatment effect is estimated as the coefficient on treatment in a Poisson
generalized linear model that does not account for treatment effect heterogeneity but does include a main effect of the covariate:

$$\log\{E(Y_i|T_i,X_i)\} = \hat{\beta}_0 + \hat{\beta}_T T_i + \hat{\beta}_X X_i$$

then we can adapt the derivations in Drake and McQuarrie\(^{18}\) (as shown in the Web Appendix) and approximate the bias of the log RR effect estimate:

$$\text{Bias}(\hat{\beta}_T) = \beta_H \left[ \bar{X}_1 - \bar{X} - (\bar{X}_1 - \bar{X}_0) \frac{p e^{\beta_T} \hat{\sigma}_1^2}{p e^{\beta_T} \hat{\sigma}_1^2 + (1 - p) \hat{\sigma}_0^2} \right]$$

where $\bar{X}$ is the overall mean of $X$, $\bar{X}_t$ and $\hat{\sigma}_t^2$ are the sample mean and variance of $X$ in patients with $T_i = t$, respectively, and $p$ is the proportion of patients with $T_i = 1$. An approximate percent bias on the unlogged scale can be found by calculating $e^{\text{Bias}(\hat{\beta}_T)} - 1$.

Since larger treatment effect heterogeneity is associated with higher estimated bias, using the maximum reasonable value for the treatment effect heterogeneity parameter will provide an upper bound for the potential bias.

If investigators would like to a priori choose a decision rule for transitioning from the pilot study to the large-scale study, we recommend focusing on the overlap metrics, since imbalance in observed covariates can generally be overcome through covariate adjustments. If a stratified analysis is planned, we recommend focusing on the proportion of patients in the region of PS overlap, with a lower proportion being less favorable and implying that a larger number of patients will likely be removed in the large-scale study. If a multivariate regression analysis is planned, we recommend focusing on estimated bias due to nonoverlap on the PS, with higher estimated bias being less favorable.
Example study

We implemented the methods discussed above in a cohort study of Cox-2 inhibitor (celecoxib, rofecoxib, or valdecoxib) use on gastrointestinal (GI) toxicity and myocardial infarctions (MI) versus ns-NSAIDs based on healthcare and prescription claims\textsuperscript{19-21}. Details are given in the Web Appendix. We mimicked a prospective pilot study design by evaluating patients sequentially, as they initiated treatment, beginning in the early post-marketing period of each Cox-2 inhibitor (1999-2005). The United States Food and Drug Administration approved Celecoxib on December 31, 1998. Rofecoxib was approved on May 20, 1999, and was pulled from the market on September 30, 2004. Valdecoxib was approved on November 20, 2001, and was pulled from the market on April 7, 2005. Covariates were created to capture known risk factors of NSAID-associated gastrototoxicity and acute MI and were based on claims in the 365 days before the index prescription.

We split the 7 years covered by the study into 14 periods of 6 months that defined 14 potential interim analyses. All analyses were performed separately using data from: 1) each of the 14 interim periods, 2) each state (all patients were enrolled in a state drug insurance program in either New Jersey or Pennsylvania), and 3) each of the 3 Cox-2 inhibitor groups versus the ns-NSAID reference group. Therefore, there were 84 potential analyses, but analyses that contained any treatment group with fewer than 100 patients were dropped, resulting in 64 analyses, all performed in R, version 2.15.2 (Vienna, Austria).

We first assessed balance across treatment groups. We estimated a PS with a logistic regression model that included linear terms for all covariates, and we used this PS for calculating the $C$-statistic. In order to estimate the percent bias due to confounding, we had to assume values for the conditional RRs that measure the associations between each
covariate and the outcomes. We estimated these associations in a group of ns-NSAID initiators from historical data (1994-1998) using outcome events recorded in the claims databases. We estimated Cox proportional hazards models including linear terms for all covariates, separately for GI events and MI events. The estimated hazard ratios from these models were used to estimate the percent bias. Although the associations between some covariates (for example, age) and treatment or outcome are likely nonlinear, modeling these effects with linear terms is often sufficient for evaluating balance and ranking strengths of association.

As a sensitivity analysis, we consulted with a rheumatologist (SYK) and asked her to classify the available covariates according to their likely association with each outcome as much less risk, slightly less risk, no association, slightly more risk, or much more risk; based on these classifications, we assigned the RR values 0.75, 0.9, 1.0, 1.11, and 1.33, respectively. In the case of continuous covariates, RRs were scaled so that these values applied to 1/6 the range of the covariate, corresponding to approximately one standard deviation for normally distributed variables. These values were chosen to produce reasonable RRs for all covariates in this example.

We also evaluated the overlap at each interim analysis. Assuming a stratified analysis in the large-scale study, we considered decision rules based on the proportion of patients in the region of overlap on the PS reaching 50%, 75%, or 90%. Assuming a regression analysis, we considered decision rules based on the estimated bias due to nonoverlap on the PS falling below 10%, 5%, or 2%. For bias estimation, we assumed RR treatment effects on GI events from 0.15 up to 0.65 ($\beta_T = -1.90$, $\beta_H = 1.47$) and on MI events from 0.80 up to 1.5 ($\beta_T = -0.22$, $\beta_H = 0.63$).
Results

Covariate balance

In each of the 64 analyses, the study size ranged from 2,620 up to 10,736 patients. Figure 2 presents the mean differences in individual covariates at each interim analysis with the accompanying legend presented in Figure 3. Because we wanted to present the results from all covariates on a single scale, we standardized differences by the within group covariate standard deviation. These plots show that in both states and for all drugs, the differences in many covariates between Cox-2 and ns-NSAID patients became smaller in magnitude over time. For example, the large over-prevalence of osteoarthritis patients and under-prevalence of black patients among Cox-2 initiators were each reduced over the study period, but remained relatively large. In contrast, a large difference in mean age was observed across the study period and did not get smaller over time.

The $C$-statistics, presented in Figure 4, also indicate that the differences between the treatment groups were reduced over time during the post-marketing period. However, at the end of the study period, the $C$-statistics were still generally greater than 0.6, representing the potential for significant bias in effect estimation without covariate control. In contrast, the estimated bias, presented in Figure 5, was not reduced over time and even appeared to increase in some circumstances. For example, the bias of estimated effects on GI events displayed a generally increasing trend across interim analyses in Pennsylvania (lower right panel). The estimates of bias were even more unpredictable when using the risk ratio values based on the expert opinion covariate classification (shown in Web Appendix Figure 2). The RR values used in these estimates are presented in Figure 3.
We suspected that the persistent differences in age and gender observed in Figure 2 were likely driving the large estimates of bias across interim analyses. To evaluate this hypothesis, we calculated estimated bias at each interim analysis controlling for age and gender (setting these RRs equal to 1.0). As shown in Web Appendix Figure 3, we found that estimated bias was generally smaller after controlling for age and gender, but it remained inconsistent from one interim analysis to the next with no clear downward trend.

Covariate overlap

In Figure 6, we present the graphical evaluation of overlap between celecoxib and ns-NSAID users for the estimated PS, age, and number of different drugs (defined by generics) for patients at the first interim analysis in New Jersey only. Although the variables presented are imbalanced between celecoxib and ns-NSAID users, their distributions are generally overlapping. This result is confirmed when evaluating the variable strata. For example, 4.1% of ns-NSAID initiators are in the highest PS decile, and 5.4% of celecoxib initiators are in the lowest PS decile. Although not shown, the analyses of the other drugs, other interim analyses, and other covariates similarly showed good overlap, except in the analyses with the smallest sample sizes; out of 176 rofecoxib initiators in New Jersey at the first interim analysis, only 4 (2.3%) were in the lowest decile of the PS.

Because treatment groups had good overlap with respect to all covariates, all decision rules considered were met relatively quickly, even under the strictest criteria. For example, the proportion of patients in the region of overlap on the PS was above 95% at all interim analyses, as shown in Web Appendix Figure 4. Therefore, under all decision rules considered for this metric, the transition to the large-scale study would have occurred after the first interim analysis (for each drug) and in both states. For the decision rules based on
estimated percent bias due to nonoverlap, shown in Figure 7, all drugs met the 10% criterion at the first interim analysis and the 5% criterion by at least the second interim analysis. However, reaching the 2% criterion took much longer, and estimated bias for the analysis of GI events in Valdecoxib initiators in Pennsylvania never fell below 2%. In general, bias due to nonoverlap was estimated to be smaller for the MI outcome than the GI outcome, due to the smaller treatment effect heterogeneity parameter used for MI.

**Discussion**

In this paper, we presented methods for evaluating the comparability of patient populations based on covariate data. These methods can be used to determine the likely bias of treatment effect estimates with and without adjustment for covariates, based on the observed covariate distributions and treatment choice. With a sequential study design that uses initial pilot data to evaluate covariates, investigators can assess the utility of restricting the patient population that is included in comparative analyses before enrolling patients in the large-scale study. In cohort studies of newly marketed medications, this approach has the advantage of potentially saving study costs by avoiding the expense of enrolling patients that should not be used for analysis. In addition, if investigators determine patient restrictions before observing outcomes, then these restrictions cannot be influenced by the result of the large-scale study.

We applied the methods introduced in this paper to an example study of Cox-2 inhibitors and ns-NSAIDS. Covariate balance was poor across all interim analyses. Therefore, comparative studies initiated at any time during this period would have required covariate adjustment. However, while the covariate mean differences and
estimated $C$-statistics generally indicated that the selectiveness of prescribing Cox-2 inhibitors was decreasing throughout the study period, the estimated bias in some scenarios indicated increasing imbalance across treatment groups, even after controlling for age and gender.

These differences may be due to the fact that estimated bias accounts for prior knowledge about the importance of each covariate to the risk of outcome, while the other metrics do not incorporate this information. We used historical data and expert opinion to quantify the magnitude of covariate associations with outcome. Estimates of bias varied across these two methods, but in each case the estimated bias was relatively large. In cases where some or all covariate associations with outcome are completely unknown, investigators will have to rely on the $C$-statistic and covariate absolute differences. These methods require fewer assumptions, but they may be more difficult to interpret.

Despite strong covariate imbalance in the example study, covariate overlap was generally good, indicating that trimmed analyses would remove few patients and regression adjusted analyses would not be subject to inappropriate extrapolation. The weakest overlap was observed at interim analyses with small study sizes. This finding demonstrates the inherent relation between study size and variable overlap; the smaller the study size, the more likely nonoverlap becomes, since sampling variability in small samples may cause low-frequency areas of the covariate distribution to be extremely sparse. Therefore, when evaluating overlap in pilot data, investigators should keep the size of the planned large-scale study in mind, as the overlap in that study may systematically differ from what is observed in pilot data if the study sizes differ. In the example study, we assumed that sizes would be similar between the interim analyses and the large-scale
study, so that the covariate overlap observed at each interim analysis was sufficient to proceed to the large-scale study. If we instead expected to enroll only 200 patients in the large-scale study, the minor nonoverlap in the tails of the PS distributions would be a concern. Therefore, we would consider excluding these patients (osteoarthritis patients with many comorbidities or co-medications), as these patients were much more likely to receive a Cox-2 inhibitor.

Although the approach demonstrated in this paper can help inform decision making on escalating a pilot study to a large-scale study with primary data collection, it is not intended to remove investigators and stakeholders from the decision. Our approach provides a quantitative and rational framework for including considerations about treatment selection in the decision making process. However, the bias estimators provided in this paper are only rough guides based on available information, and there is no threshold for adequate versus inadequate estimated bias. Reaching zero estimated bias is unlikely (even in randomized trials). External considerations may provide additional input on this decision, including considerations of the questions that the study will inform\textsuperscript{22,23}.

Observed balance and overlap can change dramatically between pilot data and the subsequent large-scale study due to systematic changes in prescriber decision making, which may be likely in the early post-marketing period, or due to simple random sampling variation if study sizes are small. Since the bias of treatment effect estimates depends on the balance and overlap in the sample in which the effects are estimated, it is paramount that investigators continue to monitor covariate distributions in the large-scale study. Finally, any confounding variables that were not measured would contribute to residual bias, and none of the methods presented in this paper are capable of assessing this bias.
References


16. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in


Figure 1: Example of a sequential cohort study of a newly marketed medication.
Figure 2: Differences in covariate means between treatment groups, standardized by the estimated within-group covariate standard deviation. Patients from each interim analysis and each state are analyzed separately. The legend is presented in Figure 3.
Figure 3: Legend identifying the color and line type for each covariate. Covariates in the legend are in decreasing order according to the standardized difference between celecoxib and ns-NSAIDs at the first interim analysis, corresponding to the ordering at the far left of the upper left plot in Figure 1, but the legend applies to all plots in Figure 1. The relative risk values used in the estimate of percent bias due to covariate imbalance in Figure 5 and Web Appendix Figures 2 and 3 are shown next to each covariate.

<table>
<thead>
<tr>
<th></th>
<th>Historical data</th>
<th>Expert opinion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MI</td>
<td>GI</td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>0.98</td>
<td>1.07</td>
</tr>
<tr>
<td>Age</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>Gastroprotective drugs</td>
<td>1.00</td>
<td>1.17</td>
</tr>
<tr>
<td>Peptic ulcer disease</td>
<td>0.90</td>
<td>1.03</td>
</tr>
<tr>
<td>Concurrent gastroprotective drugs</td>
<td>1.06</td>
<td>0.96</td>
</tr>
<tr>
<td>Number of distinct generics</td>
<td>1.02</td>
<td>1.02</td>
</tr>
<tr>
<td>Female</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>1.07</td>
<td>1.02</td>
</tr>
<tr>
<td>Number of doctor visits</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Warfarin</td>
<td>0.95</td>
<td>1.08</td>
</tr>
<tr>
<td>Oral steroids</td>
<td>0.99</td>
<td>0.91</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>1.43</td>
<td>1.05</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>1.16</td>
<td>1.15</td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td>1.21</td>
<td>1.15</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>0.91</td>
<td>0.88</td>
</tr>
<tr>
<td>ARBs</td>
<td>1.07</td>
<td>1.02</td>
</tr>
<tr>
<td>Combined comorbidity score</td>
<td>1.00</td>
<td>1.06</td>
</tr>
<tr>
<td>Clopidogrel</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td>Prior GI hemorrhage</td>
<td>1.01</td>
<td>1.44</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1.07</td>
<td>1.08</td>
</tr>
<tr>
<td>Number of prior hospitalizations</td>
<td>1.03</td>
<td>1.05</td>
</tr>
<tr>
<td>Chronic kidney disease</td>
<td>1.39</td>
<td>1.38</td>
</tr>
<tr>
<td>COPD</td>
<td>1.07</td>
<td>1.08</td>
</tr>
<tr>
<td>Number of days hospitalized</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Beta blockers</td>
<td>1.10</td>
<td>0.98</td>
</tr>
<tr>
<td>Stroke</td>
<td>1.12</td>
<td>1.15</td>
</tr>
<tr>
<td>Angina</td>
<td>1.11</td>
<td>0.89</td>
</tr>
<tr>
<td>Transient ischemic attack</td>
<td>0.97</td>
<td>1.01</td>
</tr>
<tr>
<td>Number of days in a nursing home</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>New MI</td>
<td>1.30</td>
<td>1.07</td>
</tr>
<tr>
<td>Other lipid lowering drugs</td>
<td>0.99</td>
<td>1.31</td>
</tr>
<tr>
<td>Statins</td>
<td>0.99</td>
<td>0.90</td>
</tr>
<tr>
<td>ACE inhibitors</td>
<td>1.11</td>
<td>1.08</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>1.38</td>
<td>1.15</td>
</tr>
<tr>
<td>Old MI</td>
<td>1.28</td>
<td>0.97</td>
</tr>
<tr>
<td>Other race</td>
<td>0.74</td>
<td>0.96</td>
</tr>
<tr>
<td>Diabetes drugs</td>
<td>1.38</td>
<td>1.07</td>
</tr>
<tr>
<td>Black race</td>
<td>0.71</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Figure 4: C-statistics from a PS model using all covariates, estimated separately for patients at each interim analysis and within each state. Nonselective-NSAID patients are the reference in all analyses.
Figure 5: Estimated percent bias due to imbalance on covariates for the MI outcome (top panels) and the GI outcome (bottom panels), estimated separately for patients at each interim analysis and within each state. RR values were estimated from historical data. Nonselective-NSAID patients are the reference in all analyses.
Figure 6: Densities of the estimated propensity score, age, and number of distinct drugs (defined by generics) at the first interim analysis (patients initiating treatment January 1, 1999 to June 30, 1999) in New Jersey, separately for celecoxib and ns-NSAID initiators.
Figure 7: Estimated percent bias due to nonoverlap on the PS for the MI outcome (top panels) and the GI outcome (bottom panels), estimated separately for patients at each interim analysis and within each state. Nonselective-NSAID patients are the reference in all analyses.