

Effective Health Care Program Research Reports

Number 33

Summary Variables in Observational Research: Propensity Scores and Disease Risk Scores

Patrick G. Arbogast, Ph.D.
John D. Seeger, Pharm.D., Dr.P.H.
DEcIDE Methods Center Summary Variable Working Group

Research from the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) Network



Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

May 2012

Effective Health Care Program Research Report Number 33

The DEcIDE (Developing Evidence to Inform Decisions about Effectiveness) network is part of AHRQ's Effective Health Care Program. It is a collaborative network of research centers that support the rapid development of new scientific information and analytic tools. The DEcIDE network assists health care providers, patients, and policymakers seeking unbiased information about the outcomes, clinical effectiveness, safety, and appropriateness of health care items and services, particularly prescription medications and medical devices.

This report is based on research conducted by the DEcIDE Methods Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. HHSA 290-2005-0016-I, Task Order 10). The AHRQ Task Order Officer for this project was Parivash Nourjah, Ph.D.

The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or the U.S. Department of Health and Human Services.

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials that are clearly noted in the document. Further reproduction of those copyrighted materials is prohibited without the specific permission of copyright holders.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

None of the authors has a financial interest in any of the products discussed in this report.

Suggested citation:

Arbogast PG, Seeger JD, DEcIDE Methods Center Summary Variable Working Group. Summary Variables in Observational Research: Propensity Scores and Disease Risk Scores. Effective Health Care Program Research Report No. 33. (Prepared by DEcIDE Methods Center under Contract No. HHSA 290-2005-0016-I, Task Order 10.) AHRQ Publication No. 11(12)-EHC055-EF. Rockville, MD: Agency for Healthcare Research and Quality. May 2012. <http://effectivehealthcare.ahrq.gov/reports/final.cfm>.

Summary Variables in Observational Research: Propensity Scores and Disease Risk Scores

Structured Abstract

Objectives. This paper describes the use of two types of summary scores in the context of observational research in pharmaco-epidemiology: propensity scores and disease risk scores. Either of these approaches collapses multiple potentially confounding variables into a single score and offers advantages and disadvantages. The aim is to describe best practices for creating and applying these two types of scores.

Conclusions. Settings that favor propensity scores tend to be those where there are more persons exposed to the treatment of interest than persons who have study outcomes. Another setting that favors propensity scores is when assessing a therapy's effects on multiple outcomes. Disease risk scores might be favored when assessing the effect of multiple exposures on a single outcome. Disease risk scores may also be preferable summary measures when the exposure is infrequent or consists of multiple levels and the outcome is common. Either method provides advantages for assessing treatment effect heterogeneity. A rationale for use of either summary method should be provided by the researchers who use these methods.

Contents

Background	1
Addressing Confounding	1
Confounding by Indication	1
Study Design Taxonomy.....	2
When Are Summary Variables Necessary?	3
Transparency	4
Propensity Score Development.....	5
Motivation.....	5
Good Prescribing Creates Confounding	5
Build/Estimate Propensity Score	7
Comparison Group.....	7
Selection of Covariates	7
Missing Data	7
Inclusion of Variables in the Propensity Score Model	7
Modeling.....	8
Time-Varying Characteristics	8
Interactions.....	8
Assess the Propensity Score.....	9
Propensity Score Use	10
Restriction.....	10
Stratification.....	11
Matching	12
Modeling	13
Weighting.....	13
Disease Risk Score Development	15
Motivation.....	15
Disease Risk Score Calculation	16
Disease Risk Score Assessment.....	16
Disease Risk Score Use	18
Modeling	18
Stratification.....	18
Some Unanswered Questions	19
Propensity Scores.....	19
How Best To Use the Propensity Score	19
Variable Selection	19
Disease Risk Scores	19
Matching	19
Weighting.....	19
Variable Selection	19
Recommendations.....	20
Choosing Between Propensity Score and Disease Risk Score	20
Transparency.....	20
References.....	22

Tables

Table 1. Effect of Stratification	24
Table 2. Presentation of Stratified Results.....	25
Table 3. Statins, Before Matching	26
Table 4. Statins, Effect of Matching	27
Table 5. TPA and Stroke – Comparison of Methods.....	28
Table 6. Antipsychotic Use and Risk of Sudden Cardiac Death by Cardiovascular Risk Score	29

Figures

Figure 1. Study Design Taxonomy	30
Figure 2. Hypothetical Distribution of Propensity Scores	31
Figure 3. Illustration of Propensity Scores and Non-overlap	32
Figure 4. Propensity Score Distribution Illustrating Stratification	33
Figure 5. Matched Analysis Presentation of Results	34

Appendixes

Appendix A. Bibliography	
--------------------------	--

Author affiliations:

Patrick G. Arbogast, Ph.D.¹

John D. Seeger, Pharm.D., Dr.P.H.²

DEcIDE Methods Center Summary Variable Working Group³

¹Vanderbilt University Medical Center

²Brigham and Women's Hospital and Harvard Medical School

³Contributing members of the DEcIDE Methods Center Variable Working Group:

Patrick G. Arbogast, Ph.D., Vanderbilt University Medical Center

John D. Seeger, Pharm.D., Dr.P.H., Brigham and Women's Hospital and Harvard Medical School

Becky Briesacher, Ph.D., University of Massachusetts School of Medicine

Rongwei (Rochelle) Fu, Ph.D., Oregon Health & Science University

Tobias Gerhard, Pharm.D., Ph.D., Rutgers Ernest Mario School of Pharmacy

Sean Hennessey, Pharm.D., Ph.D., University of Pennsylvania School of Medicine

Parivash Nourjah, Ph.D., Agency for Healthcare Research and Quality

Joe V. Selby, M.D., M.P.H., Kaiser Permanente

Sebastian Schneeweiss, M.D., Sc.D., Brigham and Women's Hospital, Harvard Medical School

Glen T. Schumock, Pharm.D., M.B.A., University of Illinois-Chicago

Til Stumer, M.D., M.P.H., University of North Carolina Chapel Hill

Priscilla Velentgas, Ph.D., Outcome Science Inc.

Background

This paper describes the use of two types of summary variables in the context of observational research in pharmaco-epidemiology: propensity scores and disease risk scores. Both provide a means of accounting for a large number of measured covariates through a single computed score in study design or analysis, which can be advantageous when the number of covariates to be accounted for is large relative to the number of patients or outcomes. This paper recognizes that there are numerous ways to design and conduct observational studies that may incorporate summary variables and does not seek to define how to conduct the research. Instead, the aim is to describe best practices for creating and applying these scores. Along with principles for good practice, we advocate for more transparency in the use of summary variables through providing clear and detailed descriptions of their creation, performance and use in study reports, so that readers or reviewers will be able to follow what was done and draw appropriate conclusions from the work.

Sound research depends on sound study design. We assume that the measure of exposure to treatment is accurate (i.e. the study appropriately captures people exposed to the treatment and distinguishes them from those not exposed to the treatment), and that the study's measure of outcome is accurate (study outcomes are identified with reasonable sensitivity and specificity). Further, we assume that variables that might confound the association between the treatment and the outcome have been measured. Although an important topic, a thorough discussion of unmeasured confounders is outside the scope of this paper. Because a flawed study design may not be totally remedied through analytic techniques, this paper assumes that the study design is sound and the above assumptions hold sufficiently that the primary issue to be addressed is control of confounding. Thus, the focus here will be on how to transparently use summary variable methods (propensity scores or disease risk scores) toward this end.

Addressing Confounding

If the treatment exposures of interest in an observational study were evenly distributed among people who had approximately the same risks for study outcomes, then the potential for confounding would be small. However, it is common to see differences in measures of risk between people exposed to different treatments in such studies. Accounting for these differences, which may lead to observed associations of treatment and outcome even when no causal effect of treatment exists, becomes one of the primary challenges in observational research.

Confounding by Indication

Confounding by indication is a specific type of confounding that is commonly encountered in pharmacoepidemiologic studies, and may also be referred to as channeling bias. It can be understood in a straight-forward way if the indications for therapy (reasons for choice of one treatment over another) influence the likelihood of occurrence of the outcome being studied. When treatment choice is based in whole or in part on certain patient characteristics that may be associated with study outcomes, this leads to expected differences in the occurrence of study outcomes even if the drug itself has no causal effect. Since statins are prescribed to reduce LDL levels, people who receive statins tend to have higher LDL levels than people who do not (they have the "indication" for statin therapy); and the relatively greater LDL reduction expected with rosuvastatin leads us to expect that those prescribed rosuvastatin would tend to have even higher baseline LDL levels than those prescribed other statins. An observational study of

rhabdomyolysis among rosuvastatin users relative to users of other statins would need to address selective prescribing with respect to expected LDL reduction that might mean that rosuvastatin users as a whole are at a different risk of rhabdomyolysis than users of other statins.

In some cases confounding by indication can be intractable. For example, if the indications for a therapy were clear and unambiguous and were universally followed by prescribers, then there would be no exposure variation among those with the indication for therapy. Although such conditions are rare, they might more easily occur within local settings, such as within an institution where a particular therapy is only given according to a protocol that specifies patient characteristics required for treatment. In such a circumstance, it may not be possible to find exposure variation within distinct sets of patient characteristics (i.e. everyone with a given set of characteristics is either treated or not treated). If the characteristics that determine treatment in this hypothetical example also confer a different risk of the outcome, then it may be impossible to disentangle the effect of the therapy from the effect of the characteristics that determine therapy.

Another problem with addressing confounding by indication arises when the indications are not adequately measured or available for analysis. If some patient characteristics that lead to the choice of one therapy over another are not measured and incorporated as covariates in the analytic dataset, and these same characteristics also predict outcome, then standard epidemiologic techniques may not be able to account for them (although they might be amenable to instrumental variable techniques). However, if the prescribing decision is made using patient characteristics that are measured and recorded so that they can be used analytically, and there exists some variation in exposure across a relevant range of patient characteristics, then the effect of confounding by indication can be addressed through standard epidemiologic methods, including those discussed in this paper.

Study Design Taxonomy

Pharmaco-epidemiology and epidemiology more generally address confounding either through study design, analysis or both. Study design options include restriction and matching, while analysis options include stratification, regression, and weighting. Figure 1 shows study design taxonomy.

In observational studies assessing beneficial (effectiveness) or adverse (safety) outcomes, confounding may be due to numerous comorbidities that differ according to treatment, and the potentially large number of such comorbidities complicates the use of epidemiologic techniques discussed here. All of the traditional epidemiologic methods face difficulties when accounting for numerous covariates. The design options of restriction and matching become unwieldy and eventually impossible when attempting to restrict or match on numerous variables. Restricting comparisons to subsets of the population that meet numerous conditions leads to smaller and smaller subsets available with consequent loss of ability to make the comparisons or to generalize results. Matching on numerous characteristics leads to an expansion of the number of matching categories so that finding both exposed and unexposed subjects within categories becomes impossible.

The analysis options similarly break down with an expansion in the number of covariates to account for. Stratification on numerous variables will quickly lead to many strata that do not have both exposed and unexposed subjects in them so that they do not contribute to the estimation of treatment effects. Multivariable regression encounters the more hidden problem of extrapolation where the comparisons might be made between observed data and extrapolated

data for covariate patterns where there are not both exposed and unexposed people. These extrapolated comparisons are dependent on the validity of the assumptions inherent in the modeling since they are made outside the range of analyzable data. Further, the regression model depends on adequate numbers of outcomes, which may be difficult to achieve when studying rare events. Without 8-10 events per predictor variable in the model, problems in estimation can occur, and confounder selection strategies to reduce the number of variables in the model have their own problems.¹ Weighting relies on information about the event rates within strata and thus suffers from the same problem as stratification as the number of variables to be accounted for increases.

These challenges, which arise when numerous variables need to be accounted for in either design or analysis, provide the conditions when the use of summary variables should be considered. The two main approaches, which form the topic for this paper, are to collapse variables that are predictors of exposure (propensity score), or to collapse variables that are predictors of outcome (disease risk score).

The propensity score involves collapsing variables which are predictors, or correlates, of exposure into a single summary variable.² The presence of each of the correlates of exposure confers a different probability of exposure than its absence and combinations of the variables confer a joint probability of exposure that is different than the absence of the combination. The range of values of probability of exposure is from 0 (no chance of exposure given the person's characteristics) to 1 (certainty of exposure given the person's characteristics). So, the propensity score is a value between 0 and 1 that is the predicted probability of exposure given the values of a set of variables that are correlates of exposure.

The development of the propensity score involves identifying the variables that are correlates of exposure, modeling the exposure as a function of these covariates, and estimating the probability for each individual, often using logistic regression with treatment (treatment A vs B, or treatment vs. non-treatment) as the dependent variable. Once the propensity score is developed, it can be used to address confounding through traditional epidemiologic methods (restriction, matching, stratification, modeling, and weighting).

The disease risk score approach, in contrast, involves collapsing correlates of outcome into a single summary variable (also known as confounder score, multivariate confounder score, or comorbidity score). The presence of each of the correlates of outcome confers a different probability of the outcome and combinations of the variables confer a different probability of the outcome.

Once the disease risk score is developed through a similar process of identifying variables that are correlates of outcomes, modeling the outcome as a function of these covariates, and estimating a value of the score for each individual, it is used to adjust for confounding through traditional epidemiologic methods (such as stratification or modeling).

When Are Summary Variables Necessary?

The choice of whether to use summary variables to adjust for confounding rather than traditional multivariate regression should be based in part on the ratio of expected number of covariates to adjust for to the expected number of outcomes in the study.

The rule of thumb from logistic regression modeling of being able to adjust for one variable for each 8-10 people who experience an outcome is borne out in simulation studies³ that indicate a preference for exposure-based modeling when there are few people who have the study outcome, but many who have the study exposure relative to the number of adjustment variables.

The cut-point for this preference is approximately 8 outcomes per adjustment variable. When there are more than 8 outcomes per adjustment variable, a regression model directly using predictive variables is preferred to the summary variable approach.

Transparency

A premise of this paper is that there are often several ways to arrive at a valid research result. An appreciation of the advantages and disadvantages of the summary variable methods described will assist researchers in choosing one that may be most applicable to a particular research question. Since the research methods used can affect the conclusions that may be drawn from a particular study, transparency in the use of a method allows the reader to understand and fully appreciate the work. Accordingly, transparency should be a guiding principle in the conduct and reporting of research.

The examples presented in this paper highlight applications of propensity scores and disease risk scores with the goal of identifying opportunities to use these methods to address the problem of confounding while promoting transparency in how the confounding was addressed.

Propensity Score Development

Motivation

We begin with a case example that will serve to motivate the use of propensity scores. A different example will be used to motivate disease risk scores. The cholesterol lowering drug rosuvastatin has been shown to have a larger beneficial effect on low-density lipoprotein (LDL) cholesterol reduction at a given dose than the other marketed statins.⁴ This greater relative therapeutic effect could mean that rosuvastatin also has a greater potential to produce toxicity or adverse effects of treatment. Since some adverse effects of statins, such as rhabdomyolysis, can be severe and lead to substantial morbidity, then the greater therapeutic effect of rosuvastatin would need to be weighed against this greater potential for adverse effects when prescribing rosuvastatin.

Though the question “does rosuvastatin lead to more rhabdomyolysis than other statins?” appears simple, designing a study to answer the question is complicated. The rarity of rhabdomyolysis (estimated incidence 0.4-6.0 per 10,000 person-years among statin users)⁵ means that in order to have adequate statistical power to show a two-fold difference in rates among rosuvastatin users compared to other statin users, a study would need to include tens of thousands of person-years of rosuvastatin and comparator statin exposure. This large sample size requirement means that answering the question through a randomized controlled trial would be very expensive and need to include a very large patient population followed for some years, so that an observational study might be preferred, provided that valid conclusions could be drawn. However, an observational study of rhabdomyolysis among rosuvastatin and other statin users is also complicated by a number of considerations. First, since rosuvastatin is known to have a greater therapeutic effect on LDL reduction, it will most likely be prescribed to patients with more severe cardiovascular disease and related comorbidities more often than the other statins. Second, this selective prescribing of rosuvastatin may change over time. When initially introduced to the market, rosuvastatin may be prescribed to patients with need for greater LDL reduction; however, its use will become more like that of other statins as physicians become used to prescribing it over time. Third, the epidemiology of the outcome (rhabdomyolysis) is not well understood. Since it is not known whether the cardiovascular disease risk factors associated with rosuvastatin use are related to rhabdomyolysis, it is difficult to know which of the comorbidities that might differ between rosuvastatin users and other statin users would be important to account for in the study design and analysis. To know whether an observed association of rosuvastatin use with higher rates of rhabdomyolysis compared to other statins is due to a causal effect requires addressing confounding as an alternative explanation for the association. If rosuvastatin tends to be prescribed to people who have a higher underlying risk of rhabdomyolysis, then confounding may account for some or all of any observed difference.

Good Prescribing Creates Confounding

Prescribers seek to identify appropriate candidates for a treatment, a process that may involve a “treatment/no treatment” decision and possibly a decision to identify the “best” therapy, if any, for a particular patient. Generally, good candidates for a given treatment are those in whom the benefit-risk balance is favorable, where the expectation is that the patient will have a better outcome if they receive the treatment (or the “best” treatment) than if they do not receive the treatment (or receive other less optimal treatments). This decision should incorporate all of the

known effects of the treatment, including the expected beneficial therapeutic effect as well as adverse effects of the treatment. Cost or insurance coverage can also be factored in.

The characteristics of patients who are expected to benefit are either explicitly stated in the indications for that therapy or implicitly incorporated into the prescribing decision. Explicit indications for a therapy include what is written in the product labeling including demographic characteristics (age and gender) and specifics of the condition being treated.

For example, the indications for use of rosuvastatin in the label are as follows:

CRESTOR is indicated:

(1) As an adjunct to diet to reduce elevated total cholesterol, LDL, ApoB, non-HDL-cholesterol, and triglycerides and to increase HDL in adult patients with primary hyperlipidemia or mixed dyslipidemia.

(2) As an adjunct to diet to slow the progression of atherosclerosis in adult patients as part of a treatment strategy to lower Total-C and LDL-C to target levels.

CRESTOR is contraindicated:

(1) In patients with a known hypersensitivity to any component of this product, in patients with active liver disease, which may include unexplained persistent elevations of hepatic transaminase levels, in women who are pregnant or may become pregnant, and in nursing mothers.

These indications and contraindications suggest the characteristics of patients who are likely to receive rosuvastatin (i.e., those with high LDL, low HDL, who are at risk of complications of atherosclerosis, and who do not have known hypersensitivity, liver disease, and are not pregnant). If these same characteristics also predict the study outcomes, and differ to some degree in the comparison group (i.e. patients who receive a different statin or are untreated) then confounding may result.

In addition to explicit indications for a therapy, the prescriber will also use medical judgment based on in-depth knowledge about the patient in deciding on an appropriate treatment. If a patient has high cholesterol and is on other medications, then the potential for drug-drug interactions may affect prescribing. In addition, the clinician's past experience with the treatment and patient may factor into the prescribing decision. For example, if the last patient prescribed rosuvastatin by the clinician developed rhabdomyolysis, then the clinician may be hesitant to prescribe rosuvastatin again or be highly selective in how it is prescribed. In another situation, a different therapy might be chosen for a patient who has demonstrated high levels of adherence to prior therapies, particularly if adherence to therapy can affect the safety or effectiveness of the therapy.

Other influences on the prescriber include aspects of their training and exposure to pharmaceutical representatives that might affect prescribing. Different countries or different regions within countries may have local cultures and customs as to what should be prescribed to whom and when, or be influenced by local or regional key opinion leaders. Hospital-specific guidelines for treatment and health plan formularies are more explicit regional influences on prescribing.

With regard to all of these factors which influence the choice of prescribed treatment, the concern is that a given variable may influence the choice of therapy and may also be prognostic of outcome. Such a scenario leads to confounding when the groups being compared (exposed and not exposed) also differ with respect to the prognostic variable so that the observed

association of exposure and outcome is a mix of effects (effect of the exposure and effect of the prognostic variable on the outcome).

When such a prognostic or confounding variable is measured, it may be adjusted for using standard approaches including those described in this paper. A greater concern is prognostic variables that are not measured. The use of a propensity score analysis does not address unmeasured confounding, unless the proposed unmeasured confounder is represented by other “proxy” variables within the propensity score. Sensitivity analyses can aid in defining the extent to which a hypothesized unmeasured confounder could alter the study findings.

Build/Estimate Propensity Score

Comparison Group

Once the decision to use propensity scores in study design or analysis has been made, an important first step in developing a propensity score is to define what the comparison group(s) is to be. In the rosuvastatin example, the underlying question was, “Does rosuvastatin carry with it a higher risk of rhabdomyolysis than other statins?” This question suggests that an active comparator group consisting of users of other statins would be appropriate, whereas a comparison group of untreated people with hyperlipidemia would address a different question.

Selection of Covariates

Variables to be included in the propensity score should include all of those known to contribute to confounding (correlates of both exposure and outcome). However, knowledge may be incomplete on this. A list of expected correlates of exposure can be developed based on prescribing guidelines. For example, the variables implied by the rosuvastatin indications (LDL, triglycerides, HDL, atherosclerosis) could be combined with prescribing guidelines for treatment of hypercholesterolemia (the NCEP ATP III)⁶ to arrive at an *a-priori* list of likely correlates of rosuvastatin prescribing.

However, the list developed based on expected prescribing may be incomplete. Some variables that influence the treatment decision may not be part of the explicit prescribing guidelines, and they may also be prognostic of outcomes. An empiric approach to variable identification (finding correlates of exposure through data mining techniques) can serve as a safety net as it will identify predictors of exposure that were not suspected on *a-priori* grounds.

Missing Data

Missing data within the source data set will need to be addressed in order to retain observations within the analysis. In the context of a health insurance claims database, there may be no missing data for even large numbers of subjects as most variables being considered as predictors of treatment can be defined as the presence or absence of a claim for the condition.

Inclusion of Variables in the Propensity Score Model

As noted earlier, traditional modeling constraints lead to the rule of thumb that there should be 8-10 outcome events per variable in the model. In the case of the propensity score, the outcome is treatment choice. The tradeoff will be between missing important covariates, with the possible cost of incomplete adjustment for confounding, and including too many, with the possible cost of loss of efficiency (wider confidence intervals). The inclusion of variables that

are predictive of exposure, but not predictive of outcome will lead to a loss of efficiency, and such variables should not be part of the propensity score, but this assessment depends on *a-priori* knowledge or assumptions.⁷ An approach to developing propensity scores that involves extensive empirical identification of predictors of exposure creates “high dimensional” propensity scores that will likely include some correlates of exposure that might not be associated with the outcome,⁸ so this high-dimensional propensity score algorithm includes an assessment of association between variables and outcomes.

Modeling

Standard modeling considerations apply to the development of propensity score models, and variables should be specified in ways that reflect their underlying association with exposure. For example, modeling a continuous variable as several categorical variables or as linear and quadratic terms or splines allows flexibility in the model of the association that may be superior to modeling the association using a single linear term. The model is used to compute a summary probability of receiving treatment for each individual patient given their specific covariate pattern. Logistic regression is most often used, but other approaches such as regression tree-based approaches could be used.

Time-Varying Characteristics

Time-varying characteristics need to be included in the propensity score model with values relevant to the prescribing decision. In this context, the term “time varying” applies to patient characteristics that might vary over time, but whose value is most relevant to the prescribing decision at some point prior to the prescribing decision. Accordingly, the value for the characteristic should be ascertained at or before the prescribing decision. In the rosuvastatin example, if LDL levels were available, it would be the LDL level of the patient as known to the prescriber at the time of the decision to prescribe rosuvastatin that would be important to include in the propensity score, most likely the closest in time prior to the date of the prescription. A distant pre-treatment LDL level would not be the most relevant to include, and a post-treatment LDL level could introduce bias as it would likely have been affected by the treatment.

Interactions

Interactions between the most predictive variables may need to be included in the model if the effect of one predictive variable changes according to the level of another variable. Explicit interactions with time may be needed to account for changes in the way certain variables are used in the prescribing process. As medical practice evolves over time, the weight that is applied to a specific predictor can change so that the propensity to be prescribed the drug may differ over time, even for people who possess identical covariate patterns. When the propensity score is independently developed within separate blocks of calendar time, time is implicitly accounted for in the estimated coefficients for each variable within each block of time, representing the effect of each variable on exposure in each time period.⁹ Not accounting for these time varying predictors of exposure in the development of the propensity score could lead to incomplete adjustment.

Assess the Propensity Score

The propensity score can be evaluated in a number of ways. The most direct way would be to determine how much of the confounding is removed by the use of the propensity score in analysis, however, this is generally unknown. There are however a number of indirect ways to assess the performance of the propensity score.

Since the overall aim of using the propensity score is to address confounding that might exist if directly comparing outcomes between exposed and unexposed groups, the details of the propensity score itself, such as the coefficients associated with individual variables, are generally not of interest. However, they are worth examining as a diagnostic of the propensity score model. The individual coefficients should reflect what is expected about the contribution of that variable to the treatment decision. Since LDL level is a primary indication for statin therapy and higher LDL levels are anticipated to influence the treatment choice between rosuvastatin and other statins, it is expected that a variable for LDL might have a positive coefficient within the propensity score model. Further, some variables will need refinement in the propensity score model. Is it reasonable to expect a linear association between a continuous variable (such as LDL) and the odds of initiating rosuvastatin therapy? If the individual coefficients of the propensity score do not make sense in the context of what is known or expected about prescribing of the therapy, then further assessment of the propensity score modeling process may be warranted to identify potential problems in the propensity score development that might translate into incomplete control of confounding.

An important diagnostic for the propensity score prior to its use is to examine the distribution of the propensity score among the exposed group and the unexposed group. The distribution can be viewed graphically to identify the extent of overlap in the propensity score (Figure 2). The extent of overlap has implications for the approach to analysis based on the propensity score, whether the analysis is a restriction, matching, stratification, regression adjustment, or weighting. Since each of these analytic approaches makes different assumptions, if there is little or no overlap between the propensity scores of the groups to be compared, then substantially different results might come from these different approaches. If there is extensive overlap in the distribution of the propensity scores, the similar results are likely regardless of the analytic approach with the propensity score.

An intriguing approach to using the propensity score when there is concern that some important variables are not captured in the database is propensity score calibration.¹⁰ This approach uses externally-collected data that includes the variables missing from the propensity score to adjust the propensity score as calculated without the missing variables. The resulting adjusted (or “calibrated”) propensity score is then applied to the original population providing an effect estimate that, provided the assumptions inherent in the method hold, accounts for the variables missing from the original analysis.

Another approach to account for variables missing from analysis using the propensity score is to directly obtain data on the missing variables for a sample of the source population in order to assess the effect of adjusting for them.¹¹

Propensity Score Use

Once the propensity score has been developed, a choice remains in how to use it. The propensity score can be used for restriction, stratification, matching, modeling, or weighting, and the sections that follow will address each of these.

Restriction

A fundamental observational research tool, restriction, can be combined with propensity score analyses. One application of restriction might involve examining the distribution of the propensity score among exposed and non-exposed subjects. The aim is to identify potential regions of non-overlap at the tails of the distribution (subjects from one group who possess extreme high or low propensity score values with no corresponding subjects in the other group). Comparative analyses that incorporate the non-overlapping subjects in the tails would be extrapolating from a region of observed data to a region without data, and the extrapolation could lead to erroneous conclusions if the functional form of the relationships among variables changes outside the observed range. A prudent approach is to trim the data (excluding the non-overlap in the tails of the distribution) in order to ensure overlap in covariates. Figure 3 is an illustration of propensity scores and non-overlap.

Restriction in this way shares some features of matching exposed subjects to unexposed subjects on the basis of similar propensity scores, in that matching generally has the effect of removing subjects in the tails from one group when there are not subjects in the other group with comparable propensity scores.

An additional argument in favor of excluding the tails of the propensity score distribution is that there is likely to be the greatest unmeasured confounding in those subjects who are treated against expectation. The unknown characteristics of the subjects in the tails of the propensity score distribution may be unmeasured confounding or may be effect modification, a distinction that may be difficult to make.¹² Relatedly, non-overlapping regions of the propensity score may include patients who have contraindications to one of the treatment options, or individuals who have other medical characteristics which lead them to be undesirable candidates for one of the treatment options. Since they are not in fact candidates for one of the treatments, comparisons of treatment effectiveness and safety may not be generalizable or relevant to these individuals.

A different application of restriction is to focus attention on a subset of the population that might be at risk for medication dispensing errors. Medication dispensing errors, which occur when the wrong drug is dispensed to a patient, can be thought of as situations where a person is given a drug that does not match that person's conditions (demographics, comorbidities, etc.). This situation is what happens in the tails of the propensity score distribution. At one end, people who have all the right indications for a treatment (and correspondingly high propensity scores) fail to receive the treatment. At the other end, people who do not have conditions that are common among recipients of the treatment and do not otherwise appear to be appropriate candidates for the treatment (low propensity scores) receive the treatment. When observed, these situations could represent dispensing errors (i.e., the pharmacy dispensed one medication when a different medication was intended), and the subset of the population so identified is the one of interest for more detailed analyses.

Such an application was used in a study of potential medication name confusion errors between the medications Amaryl (glimepiride) and Reminyl (galantamine). All dispensings of the two drugs were identified, and subjects who received glimepiride who did not have a

diagnosis of diabetes were examined more closely. For a subset, medical records from an office visit around the time of the dispensing were sought to determine which medication might have been intended. Although a pilot study using this approach identified no apparent name recognition errors in dispensings identified, the method demonstrated that it could be used, and the study allowing for an estimated upper limit for the dispensing error to be made.¹³

Advantages of restriction include greater transparency and increased validity of comparisons. Disadvantages of restriction involve the exclusion of some exposed people. If a beneficial or adverse effect occurs differently among the set of excluded people (those to whom the restriction was not made) then this form of effect measure modification will not be observed.

Stratification

Another approach to the use of propensity scores is stratification. Patients who receive a given therapy and their comparators can be divided into subsets according to levels of a third variable as a covariate. By examining the association between treatment exposure and outcome risk within subsets of patients with similar covariate levels, the covariate is unable to produce confounding since those being compared have similar values of the covariate. The stratification of comparisons by a single covariate can be extended to several covariates so that with two covariates, four subsets are created (one group who possess both covariates, two groups who possess one of the two covariates, and one group who possess neither of the covariates). As the number of covariates increases, so does the number of strata, but at a considerably faster rate so that the approach becomes unwieldy with numerous covariates.

Accordingly, the numerous covariates can be collapsed into a single value (the propensity score), allowing for stratification that is not unwieldy, but nevertheless accounts for the numerous covariates. Stratification can be applied to continuous variables such as the propensity score by dividing the data into categories that correspond to ranges of the continuous variable. Because of the potential for residual confounding remaining due to variability of the propensity score within each strata, this process leads to a tradeoff with the simplicity of fewer strata balanced against more complete removal of confounding with more strata. The extent of confounding removed can be estimated based on the number of strata,¹⁴ and five strata are generally considered enough to remove approximately 90% of the confounding due to the variable. Figure 4 is an example of propensity score distribution.

When presenting an analysis that uses stratification by levels of the propensity score, transparency would suggest that numerous tables are needed to show the balance between comparison groups within strata. The balance between treated and comparison groups within each stratum might be shown in order to provide the reader of the research with a sense of comfort that the stratification on the propensity score actually addressed the potential confounding that might be present in the crude analysis.¹⁵ Table 1 shows the effect of stratification, and Table 2 is a presentation of stratified results.

Stratification can also be used for assessing consistency of effect across strata.¹² An analysis stratified on the propensity score can be useful in conjunction with other analytic techniques as it can show whether the effect of treatment is consistent across strata of the propensity score. This may be useful as a diagnostic of the balance achieved by the propensity score as a change in treatment effect could suggest unmeasured confounding. It may also be useful to identify subgroups where the effect of the drug differs. For example, perhaps an adverse effect of a drug is only apparent when the drug is prescribed to people who really do not have the indication for the drug (possibly reflecting off-label use) and have low propensity scores.

An advantage of stratification is transparency in that balance on covariates achieved through use of the propensity score can be shown explicitly when using stratification. A further advantage is that many readers of the research result will either be familiar with the technique of stratification or find it easy to understand so they can follow what was done and be able to better interpret the results of the analysis.

A disadvantage of stratification is that in order to be transparent many tables may be required, making for a potentially unwieldy presentation. Additionally, residual confounding within strata may cause bias.^{16,17}

Matching

Matching on the propensity score offers an intuitive approach to making comparisons. For each person who receives the therapy, a person who does not receive the therapy with a similar propensity score is selected and their follow-up is compared. Matching on the propensity score as a single variable has the effect of matching on all of the components of the propensity score, without the drawback of matching on numerous individual variables, which leads to greater and greater difficulty in finding appropriate matches due to the expansion in the number of potential matching categories. Intuitively, the follow-up of these subjects should not be confounded by any of the components of the propensity score, so that a straight-forward comparison of outcomes observed in the follow-up will not be confounded and further analysis to account for the variables may not be needed.

Table 3 shows statistics on statins before matching. Table 4 shows the effect of matching on the statin statistics. Figure 5 is a presentation of the results of matched analysis.

Any matching algorithm can be used to identify and retain treated and comparator subjects who have similar propensity scores, and they generally fall into one of two categories. These are fixed or variable caliper approaches. In a fixed caliper approach, potential comparators are identified within a fixed caliper around the propensity score of each treated subject. Then the match is chosen either through a random process or by identifying the closest of the potential matches. Variable caliper matching identifies potential matches within an initially narrow caliper and progressively extends the caliper if no match is found.¹⁸

Matching can be performed in a ratio other than 1:1. In the rosuvastatin study,⁴ a matching ratio of up to 1:4 was used to improve study power by increasing the size of the comparator group (the treated group was limited by the number of rosuvastatin initiators within the data source).

Matching can be used in combination with stratification to examine heterogeneity of effect and potential effect measure modification.

The advantages of matching include transparency in that once matched, the groups to be compared can be explicitly described so that any reviewer can see whether the characteristics of the compared groups might be subject to confounding, by presenting a comparison of characteristics similar to a “Table 1” from a clinical trial. The intuitive appeal of matching is another advantage in that the direct comparison of two groups forms the foundation of many research studies, making propensity score matched studies “look” like other studies that may be familiar to readers. Matching, which of necessity also restricts the comparisons to the regions of the distributions with overlapping propensity scores confines the analysis to the most relevant part of the population (those patients for whom the treatment decision could go in one direction or the other: empirically people with the same set of characteristics are sometimes treated and sometimes untreated).

The primary disadvantage of matching is the exclusion of unmatched subjects, which will be mainly exposed subjects that appear in the tails of the PS distribution. This problem is more likely in the upper end of the propensity score distribution where there exist exposed subjects with high propensity scores and no or few unexposed subjects with correspondingly high propensity scores, and is more likely to occur when there is less overlap between the distributions of the propensity score between groups, and/or when there is not a large pool of comparators from which to draw. This may limit the generalizability of the study's findings. The corresponding inability to directly address the effect of treatment in the tails of the propensity score distribution can be considered a limitation of matching, but stems from a lack of directly comparable subjects that is being made explicit by the analytic method.

Modeling

The propensity score can be included as a covariate in a model that otherwise includes only treatment as an independent variable and outcome as a dependent variable. The inclusion of the propensity score as a single summary variable will adjust for confounding to a similar degree as would inclusion of the numerous component variables of the propensity score in the model, but it consumes fewer degrees of freedom. Of course, this gain in efficiency depends on a number of assumptions about how the propensity score was constructed. Table 5 is a comparison of methods used for TPA and stroke.

A theoretical concern in the use of the propensity score as a term in a model is that the removal of confounding depends on the association between the propensity score and the outcome being correctly specified in the model. Since the propensity score includes many variables, the shape of the association between the propensity score and the outcome may be unpredictable and prior literature is unlikely to help (the association between the collection of variables in the propensity score and the outcome has never been described). Although a theoretical concern, the magnitude of residual confounding from an incorrectly specified propensity score in an outcome model may not be enough to qualitatively change the result. Further, the propensity score can be modeled with polynomial terms or splines to improve the fit so that it more closely approximates the underlying relationship between the propensity score and outcome.

An advantage of using the propensity score as an adjustment variable within the context of an outcome model is that such a use may be familiar for many readers of the research making it accessible to them. Further, it has the advantage of using all of the subjects with the treatment of interest and their comparators (subject perhaps to trimming areas of non-overlap in the tails through restriction if desired).

A disadvantage of the approach is that transparency may be reduced by inclusion of the propensity score to a single term in an outcome model rather than offering the opportunity to observe the individual coefficients for the various component variables as in a traditional multivariate analysis. Also, without examining and removing the areas of non-overlap in the propensity score distributions through restriction, some multivariable extrapolation may occur as mentioned above.

Weighting

Weighting by the propensity score addresses confounding by reweighting the treated group, the comparator group, or both so that the compared groups have a similar propensity score distribution and comparisons can be made between them that are unconfounded by the

components of the propensity score. There are several considerations to be made when using weighting on the propensity score. First, the standard population to which the treated and comparator groups will be weighted must be chosen. Generally the choice is between using the entire population (treated and comparators) as the standard population (Inverse Probability of Treatment Weighting [IPTW] weighting) or using the treated group as the standard population through standardized morbidity ratio (SMR) weighting. This choice has implications for the analysis, particularly in settings where there is effect measure modification.¹² In the presence of effect-measure modification, particularly in the region of a sparsely-populated tail of the propensity score distribution that differs between treated and untreated individuals, the IPTW weighting approach can lead to extremely large weights being applied to a small number of individuals. The large weight these individuals take on can lead to inferences that, while perhaps appropriate for one causal question (e.g., a comparison of treating the entire population to not treating the entire population) may not be appropriate for a different causal question (e.g., a comparison of people who actually received the treatment to similar, but untreated people).

A primary advantage of weighting is that all subjects in an analysis can be used so that there may be less concern about excluded subjects.

A disadvantage is the loss of transparency that accompanies the use of weighting. A tabular presentation of the characteristics of the compared groups is not possible in the same way as with matching or stratification. Further, the lack of familiarity with this approach may lead to misunderstandings about the appropriate standard population to use and incorrect inferences can follow.

Disease Risk Score Development

Motivation

Automated databases are increasingly used in pharmacoepidemiologic studies. These databases include records of prescribed medications and encounters with medical care providers from which one can construct very detailed surrogate measures for both drug exposure and covariates that are potential confounders. Often it is possible to track day-by-day changes in these variables, if this is appropriate for the question under study. However, while this information is often critical for study success, the potentially very large number of medications and comorbidities to be accounted for can pose challenges of how best to incorporate many covariates in the statistical analysis, when investigating the association between study outcomes and exposure of interest.

The following hypothetical motivating example, based on studies of non-steroidal anti-inflammatory drugs (NSAIDs) and cardiovascular disease, illustrates some of these challenges. It assumes a cohort study of NSAIDs and cardiovascular disease performed in an automated health care database, with both an exposure and covariates that can change on a daily basis throughout follow-up. Because the effect of NSAIDs on cardiovascular risk is thought to be acute, the primary comparison of interest is current use of NSAIDs versus nonuse, but there also is a recent use category to reduce misclassification due to intermittent use and former use categories that provides an assessment of confounding by indication.

Because data suggest that NSAID cardiovascular effects vary according the specific drug,^{19,20,21} it is now questionable to conduct analyses pooling data across individual medications in this class. In this example, we assume that there are nine individual drugs of interest, six traditional NSAIDs and three newer selective inhibitors of COX2, or coxibs. We also assume that exposure will be further classified according to three dose categories, as for some drugs there is evidence for dose-response. This specification of the study question thus leads to a categorical exposure variable with 30 levels (9*3 levels for current use, 1 for intermittent use, 1 for former use, and 1 for non-use).

The study defines surrogate measures of cardiovascular risk factors from medical care encounters. For example, a prescription for a lipid lowering agent or a hospitalization with a diagnosis of angina pectoris will be considered as surrogates for hyperlipidemia or clinically important coronary artery disease. It is not uncommon to thus identify 50 or more covariates. For example we may identify 20 medications, 10 diagnostic groups for hospital discharges, 10 diagnostic groups for outpatient visits, and 10 indicators of other diseases (such as rheumatoid arthritis or chronic obstructive pulmonary disease) that might plausibly affect cardiovascular risk. We assume in this example that each of the 50 variables has only two levels, present or absent, although in practice there often are more levels (e.g., taking into account recent hospitalization).

This type of study poses several challenges for statistical analysis. The first is computational complexity. Even with modern computing capacity, fitting regression models with datasets consisting of tens-to-hundreds of thousands of patients and 80 time-dependent covariates will require substantial computing power, and analyses may be cumbersome, particularly if there are dependencies (e.g., allowing an individual to enter the cohort multiple times) that require more complex variance estimation. This in turn may inhibit performance of important sensitivity analyses. Furthermore, if the number of disease cases is small, parameter estimates based on

large sample theory, such as those provided by widely used regression programs, may be inappropriate.

Second is the question of model specification. Because some of the covariates may not be associated with increased cardiovascular risk, variable selection procedures may be considered to improve exposure estimate precision and reduce computational complexity. However, this may involve subjective decisions such as the type of variable selection procedure, whether to base selection on p-values or change in exposure parameter estimates, and the numeric cutoffs (e.g., $p=0.05$, 0.10, 0.20) for variable inclusion. Because many of the risk factors for cardiovascular disease confer relatively modest increases in disease probability, variable selection procedures may exclude important covariates from the final model. Furthermore, techniques for constructing parsimonious models, such as stepwise regression, have limitations that can lead to underestimation of standard errors for exposure estimates.²²

A third challenge is the question of effect modification. Clinicians are interested in whether or not the risk conferred by an NSAID varies according to baseline cardiovascular risk status. However, with 50 variables that measure this factor, the effect modification analyses may be cumbersome.

An approach to handle this is to construct a disease risk score. The term disease risk score could be replaced with a more generalized term of outcome propensity score to account for situations where the outcome of interest might not be a disease at all, but rather an effectiveness outcome. However, the term ‘disease risk score’ will be used throughout this paper. The disease risk score is analogous to the propensity score in that it calculates a summary measure from the covariates. However, the disease risk score estimates the probability or rate of disease (or outcome) occurrence conditional on being unexposed for all members of the study cohort, regardless of their true exposure status. The association between exposure and disease (or outcome) is then estimated adjusting for the disease risk score in place of the individual covariates.

Disease Risk Score Calculation

Given that the disease risk score estimates the probability of disease occurrence in the absence of the exposure, the following method for calculating this score is proposed. First, an appropriate model linking the covariates to the outcome is selected (e.g., logistic regression for a binary outcome with fixed follow-up). Covariates are typically selected based on *a priori* clinical knowledge of their suspected association with exposure and outcome, similar to constructing propensity score models (and many of the same consideration apply). This is then used to fit a model in the unexposed group from which one can estimate the probability of disease occurrence. This estimated probability, calculated for each member of the cohort under the assumption of no exposure, is the estimated disease risk score.

Disease Risk Score Assessment

The disease risk score is often categorized into groups such that the lowest group consists of patients with no cardiovascular risk factors (in the case of the association between NSAIDs and CV disease) and the remaining groups are corresponding percentiles (e.g., quintiles, deciles) among patients with at least one risk factor. Regression models are then fit relating this categorized risk score to the outcome with the lowest risk score group as the referent. Then the odds ratio (OR) (or hazards ratio [HR] or relative risk [RR]) for the other risk score categories are assessed to see if they are reasonable. If any are questionable (e.g., an OR of 50), then the

disease risk score model is re-examined to assess the problem. Otherwise, it is used in the regression analysis.

Disease Risk Score Use

Modeling

The disease risk score can be included as a covariate in the regression model relating the exposure to the outcome in place of the individual covariates used to derive the risk score. This will substantially reduce the degrees of freedom in the regression model. The disease risk score can be entered as a continuous variable; however, score quantiles are typically used (e.g., quintiles, deciles) because in practical applications the risk score often is not linearly related to exposure or outcome. Further, in moderate risk populations when the covariates are determined from medical care encounters, a substantial fraction of the cohort may have none of the medical care encounters used to define the covariates and thus will have the same estimated risk score. This group can serve as the referent from which disease risk will be estimated for the other risk score groups.

Stratification

As with propensity scores, disease risk scores can also be used for stratified analyses. As mentioned above, the disease risk score is often categorized into groups in which the lowest group consists of cohort members in the lowest risk stratum (in this case, with no cardiovascular risk factors), and the remaining groups are corresponding percentiles among cohort members with at least one risk factor. Each stratum would consist of cohort members with the same or similar disease risk. For instance, the lowest stratum would consist of cohort members with no disease risk factors. The same advantages and disadvantages of stratification by propensity score apply to disease risk scores.

A common question in pharmacoepidemiologic studies of drug safety is whether the risk conferred by a drug varies according to patients' baseline risk for the outcome under study. In the NSAID example, this concept might be expressed as a concern that patients with a high underlying risk of cardiovascular disease will exhibit a larger than average response to the adverse cardiovascular effects of these agents.

The disease risk score provides a natural way to examine this type of effect modification. For example, in a study of antipsychotics and sudden cardiac death, cohort members were grouped according to either having no cardiovascular risk factors or tertiles of the cardiovascular risk score if they possessed at least one risk factor. Rate-ratios associated with antipsychotic therapy were calculated for each group, noting that risk seemed greatest for the highest cardiovascular risk factor group.²³ Specifically, in cohort members with mild, moderate, or severe cardiovascular disease or no disease, the incidence of sudden cardiac death among current moderate-dose antipsychotic users was at least 60% greater than that for comparable nonusers, with rate ratios of 1.60 (Table 6). When there are a large number of covariates that influence disease (or other outcome) risk, in this case cardiovascular risk, this type of analysis would be difficult to perform without some summary measure of that risk.

Some Unanswered Questions

Propensity Scores

How Best To Use the Propensity Score

Once a propensity score has been built, the most appropriate use of the propensity score (restriction, stratification, matching, modeling, or weighting) can be unclear. Additional research and methodologic clarification would be of value to assist researchers when deciding how to apply the propensity score.

Variable Selection

The selection of variables to be included in a propensity score is an area where additional guidance would be welcome. The tradeoff between inclusion of potentially confounding variables and their exclusion would be often faced with greater confidence if informed by additional methods development.

Disease Risk Scores

Matching

Matching methods have been developed, studied, and applied for propensity scores. For study designs such as case-control studies, matching on disease risk score may be desirable and possibly preferable to matching on propensity score. Future research in the area of matching with disease risk scores would be useful.

Weighting

For propensity scores, weighting methods have been developed and applied. Specifically, inverse-probability of treatment weighting and standardized mortality ratio weighting have been investigated.^{24,25} Future research should include examining similar weighting methods for disease risk scores.

Variable Selection

Variable selection has been investigated for propensity scores.⁷ Specifically, bias, variability, and mean squared error had been examined when the propensity score includes true confounders, covariates related only to the exposure, and covariates related only to the outcome. Future research should include examining variable selection for disease risk scores.

Recommendations

Choosing Between Propensity Score and Disease Risk Score

Settings that favor propensity scores tend to be those where more persons exposed to drug of interest than have study outcomes (i.e. a common exposure and a rare outcome). Another setting that favors propensity scores is when there might be interest in examining the effect of a therapy on multiple outcomes. In such a setting, considerable efficiency could be gained by using the propensity score to build matched cohorts whose follow-up is unconfounded by numerous variables, simplifying analysis across the multiple outcomes. Disease risk scores might be favored for study of research questions involving multiple exposures and a single outcome. When the exposure is infrequent or consists of multiple levels and the outcome is common, it may be a preferable summary measure to use. Also, when there is treatment effect heterogeneity across several medications and comorbidities, such as those observed for antipsychotic use and risk of sudden cardiac death across levels of cardiovascular disease risk, disease risk scores may be preferable.

A rationale for use of either summary method (propensity scores or disease risk scores) should be provided by the researchers who use these methods. Either the rarity of the outcome in the face of numerous plausible confounders, or the rarity of the exposure with numerous confounders are reasonable rationales. Other sound reasons could include construction of a single set of cohorts to follow for multiple outcomes (for propensity scores), or a single set of outcomes for which to evaluate the effects of multiple exposures (for disease risk scores). Another rationale might be that the propensity score or disease risk score offers transparency not available from other analytic approaches.

When the propensity score is used, the explicit choice of the comparison group(s) to address the research question should be addressed. The availability of key variables likely to be confounders (either within the data source directly or by proxy, or obtainable as a supplement to the database) is also an important question to address. The variable selection for the propensity score should be explicitly described so that readers may understand the process that led to some variables being included and others being excluded from the propensity score. The propensity score should have some diagnostics performed, such as a comparison of the distribution among treated people and comparators. The propensity score should be used in a way that is appropriate to answering the scientific question.

For disease risk scores, when the exposure is infrequent or consists of multiple levels and the outcome is common, this may be the preferable summary measure to use. Also, when there is treatment heterogeneity across several medications and comorbidities, such as those observed assessing antipsychotic use and risk of sudden cardiac death across levels of cardiovascular disease risk, disease risk scores may be preferable.

Transparency

Finally, transparency regarding the approaches used to incorporation of summary variables in study design and analysis is important so that readers of the work can assess what has been done and draw conclusions regarding the study findings and potential non-causal explanations. We suggest clear descriptions and characterization of the study population using numbers (rather than just percentages) to facilitate readers' use of the report by making explicit the population to which inferences are to be made. Tabulations of analytic results in formats corresponding to the

analyses performed (e.g. restriction, stratification, matching, etc.) in accordance with some of the examples provided in this report will also contribute to the clarity and interpretability of the study.

References

1. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol* 2008;167:523-9.
2. Joffe MM, Rosenbaum PR. Propensity scores. *Am J Epidemiol* 1999;150:327-33.
3. Cepeda MS, Boston R, Farrar JT, et al. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* 2003;158:280-7.
4. McAfee AT, Ming EE, Seeger JD, et al. The comparative safety of rosuvastatin: a retrospective matched cohort study in over 48,000 initiators of statin therapy. *Pharmacoepidemiol Drug Saf* 2006;15(7):444-53.
5. Graham DJ, Staffa JA, Shatin D, et al. Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. *JAMA* 2004;292:2585-90.
6. Expert Panel on the Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA* 2001;285:2486-97.
7. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol* 2006;163:1149-56.
8. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiol* 2009;20:512-22.
9. Seeger JD, Kurth T, Walker AM. Use of propensity score technique to account for exposure-related covariates. An example and lesson. *Med Care* 2007;45(10):S143-8.
10. Sturmer T, Schneeweiss S, Avorn J, et al. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol* 2005;162:279-89.
11. Eng PM, Seeger JD, Loughlin J, et al. Supplementary data collection with case-cohort analysis to address potential confounding in a cohort study of thromboembolism in oral contraceptive initiators matched on claims-based propensity scores. *Pharmacoepidemiol Drug Saf* 2008;17:297-305.
12. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol* 2006;163(3):262-70.
13. Cole JA, Zhu S, Russo LJ, et al. Use of propensity scores to identify possible medication dispensing errors among patients with Alzheimer's disease. *Pharmacoepidemiol Drug Saf* 2005;14:S21.
14. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968;24:295-313.
15. Eng PM, Ziyadeh N, Nordstrom BL, et al. Incidence of selected outcomes among matched cohorts of initiators of racemic zopiclone, temazepam, and zolpidem in the General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2005;14:S22-3.
16. Austin PC, Grootendorst P, Normand ST, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine* 2007;26:734-753.
17. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine* 2007; 26: 3078-3094.
18. Parsons L. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. Available at: <http://www2.sas.com/proceedings/sugi26/p214-26.pdf>.
19. Ray WA, Stein CM, Hall K, et al. Non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease: an observational cohort study. *Lancet* 2002;359:118-123.

20. Ray WA, Stein CM, Daugherty JR, et al. COX-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. *Lancet* 2002;360:1071-73.
21. Graham DJ, Campen D, Hui R, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet* 2005;365:475-81.
22. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Med* 1989;8:771-83.
23. Ray WA, Meredith S, Thapa PB, et al. Antipsychotics and the risk of sudden cardiac death. *Arch Gen Psychiatry* 2001;58:1161-67.
24. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics* 1992;48:479-95.
25. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology* 2003;14:680-6.

Tables

Table 1. Effect of stratification

Propensity Score			Propensity Score		
Quintile	Subjects	Person-Years	Quintile Limits	Mean	Min, Max
All Quintiles					
Temazepam	93,011	280,712	----	0.31	0.02, 0.97
Zopiclone	54,592	126,002	----	0.47	0.03, 0.97
Quintile 1					
Temazepam	26,957	107,774	0, 0.17	0.12	0.02, 0.17
Zopiclone	2,563	10,409	0, 0.17	0.13	0.03, 0.17
Quintile 2					
Temazepam	22,402	79,934	0.17, 0.30	0.23	0.17, 0.30
Zopiclone	7,119	25,645	0.17, 0.30	0.24	0.17, 0.30
Quintile 3					
Temazepam	17,833	47,472	0.30, 0.42	0.36	0.30, 0.42
Zopiclone	11,687	31,988	0.30, 0.42	0.37	0.30, 0.42
Quintile 4					
Temazepam	14,784	29,458	0.42, 0.55	0.48	0.42, 0.55
Zopiclone	14,737	30,325	0.42, 0.55	0.49	0.42, 0.55
Quintile 5					
Temazepam	11,035	16,073	0.55, 1.00	0.64	0.55, 0.97
Zopiclone	18,486	27,635	0.55, 1.00	0.66	0.55, 0.97

Table 2. Presentation of stratified results

Propensity Score Quintile	95% CI		
	Relative Risk	Lower	Upper
All Quintiles	Crude		
Effect: Tem-Zop	1.01	0.84	1.21
Quintile 1	Quintile-Specific		
Effect: Tem-Zop	1.03	0.59	1.77
Quintile 2			
Effect: Tem-Zop	0.92	0.63	1.34
Quintile 3			
Effect: Tem-Zop	1.13	0.78	1.63
Quintile 4			
Effect: Tem-Zop	0.85	0.55	1.31
Quintile 5			
Effect: Tem-Zop	1.63	0.96	2.77
All Quintiles	Adjusted for Quintile of Propensity Score		
Effect: Tem-Zop	1.05	0.86	1.27
All Quintiles	Adjusted for Continuous Propensity Score		
Effect: Tem-Zop	1.06	0.87	1.29

CI = confidence interval

Table 3. Statins, before matching

Variable	Initiators N=4144	Non-Initiators N=4144	P-Value
Lipid-related labs	26.04	13.58	<0.0001
Different prescription drugs	5.02	2.88	<0.0001
LDL level (mg/dL)	180.66	155.08	<0.0001
Triglyceride level (mg/dL)	202.66	166.91	<0.0001
Cardiovascular-related prescription drugs	0.59	0.26	<0.0001
Cardiovascular-related Visits	1.11	0.27	<0.0001
Age (years)	62.04	58.02	<0.0001
Physician visits	7.69	6.31	<0.0001
Ischemic heart disease	20.27%	5.57%	<0.0001
HDL level (mg/dL)	43.29	46.60	<0.0001
Cardiovascular-related diagnoses	0.28	0.12	<0.0001
Cardiovascular-related hospitalizations	0.56	0.13	<0.0001
MI	11.58%	2.92%	<0.0001
Angina	11.92%	3.14%	<0.0001
Unstable angina	10.59%	2.22%	<0.0001
Smoking	25.80%	18.10%	<0.0001
Hypertension	19.96%	12.96%	<0.0001
Labs	10.48%	10.74	0.0074
Hospitalizations	0.22	0.08	<0.0001
Male	53.14%	47.61%	<0.0001

HDL = high-density lipoprotein; LDL = low-density lipoprotein; MI = myocardial infarction

Table 4. Statins, effect of matching

Variable	Matched at 0.01 Propensity Score		
	Initiators N=2901	Non-Initiators N=2901	P-Value
Lipid-related labs	24.90	24.64	0.4987
Different prescription drugs	4.57	4.54	0.7639
LDL level (mg/dL)	177.84	177.58	0.7837
Triglyceride level (mg/dL)	200.34	200.50	0.9626
Cardiovascular-related prescription drugs	0.51	0.51	0.9367
Cardiovascular-related visits	0.74	0.83	0.1249
Age (years)	61.47	61.68	0.5030
Physician visits	7.25	7.27	0.8732
Ischemic heart disease	15.13%	15.48%	0.7428
HDL level (mg/dL)	43.51	43.55	0.9079
Cardiovascular-related diagnoses	0.21	0.23	0.2145
Cardiovascular-related hospitalizations	0.40	0.39	0.7929
MI	7.89%	8.69%	0.3169
Angina	8.51%	8.72%	0.8151
Unstable angina	7.14%	7.31%	0.8393
Smoking	23.85%	24.27%	0.7355
Hypertension	16.58%	17.99%	0.1649
Labs	10.45	10.48	0.7874
Hospitalizations	0.16	0.16	0.6915
Male	52.33%	52.09%	0.8747

HDL = high-density lipoprotein; LDL = low-density lipoprotein; MI = myocardial infarction

Table 5. TPA and stroke – comparison of methods

Method	No.	OR	95% CI
Crude model	6,269	3.35	2.28, 4.91
Multivariate model*	6,269	1.93	1.22, 3.06
Matched on propensity score	406	1.17	0.68, 2.00
Regression adjusted with propensity score	6,269	1.53	0.95, 2.48
Propensity score, continuous multivariable*	6,269	1.85	1.13, 3.03
Propensity score, deciles multivariable*	6,269	1.76	1.13, 2.72
Weighted models	6,269	1.96	1.20, 3.20
IPTW	6,269	10.77	2.47, 47.04
SMR weighted	6,269	1.11	0.67, 1.84

CI = confidence interval; IPTW = inverse probability of treatment weighted; OR = odds ratio; SMR = standardized mortality ratio

* Adjusted for age, gender, time from symptoms to hospital admission, Rankin scale, paresis, aphasia, state of consciousness, transportation to the hospital, admitting ward, admitting hospital, history of hypertension, diabetes, atrial fibrillation, other cardiac illnesses, previous history of stroke, and interaction terms for follow-up time and age, time from symptoms to admission to the hospital, and Rankin scale.

Source: Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of non-uniform effect. *Am J Epidemiol* 2006 June 24;163(3):262-70. ©Oxford University Press, 2006. Used with permission.

Table 6. Antipsychotic use and risk of sudden cardiac death by cardiovascular risk score

Cardiovascular Disease	IRR (95% CI)*	Excess SCDs[†]
None	1.60 (0.89 – 2.87)	4
Mild	3.18 (1.95 – 5.16)	21
Moderate	2.12 (1.08 – 4.14)	23
Severe	3.53 (1.66 – 7.51)	367

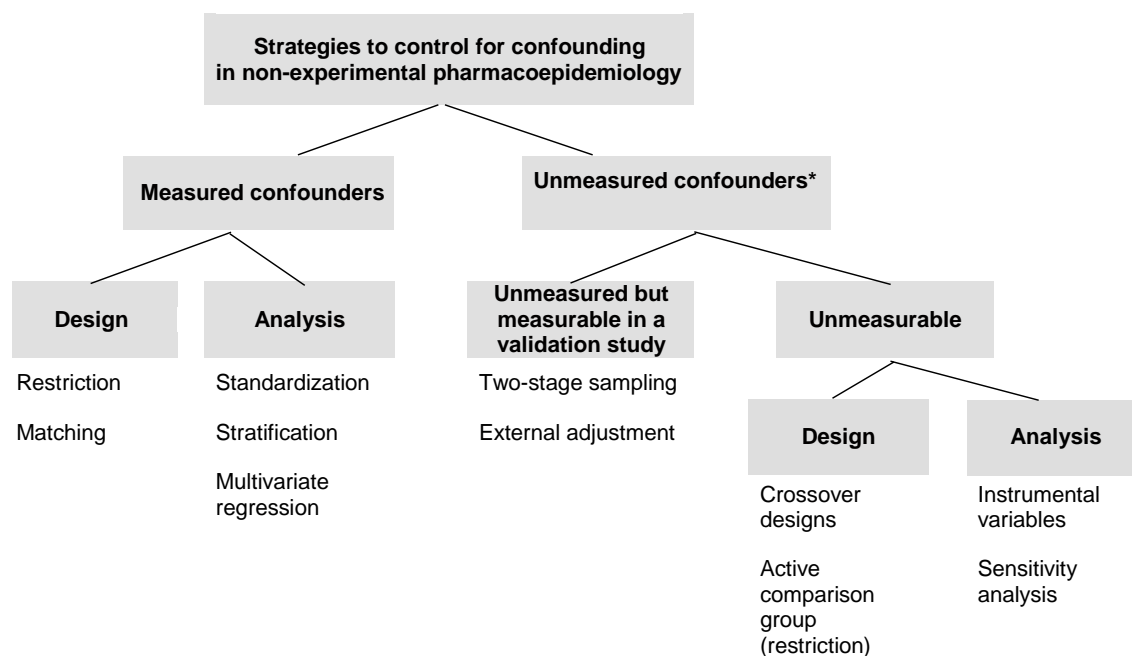
CI = confidence interval; IRR = incidence rate ratio; SCD = sudden cardiac deaths

* Among current moderate-dose antipsychotic users with nonusers as the referent.

[†] Rates are per 10000 person-years.

Figures

Figure 1. Study design taxonomy



Source: John Wiley & Sons, Inc., 2006. Used with permission. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf* 2006; 15(5):291–303.

Figure 2. Hypothetical distribution of propensity scores

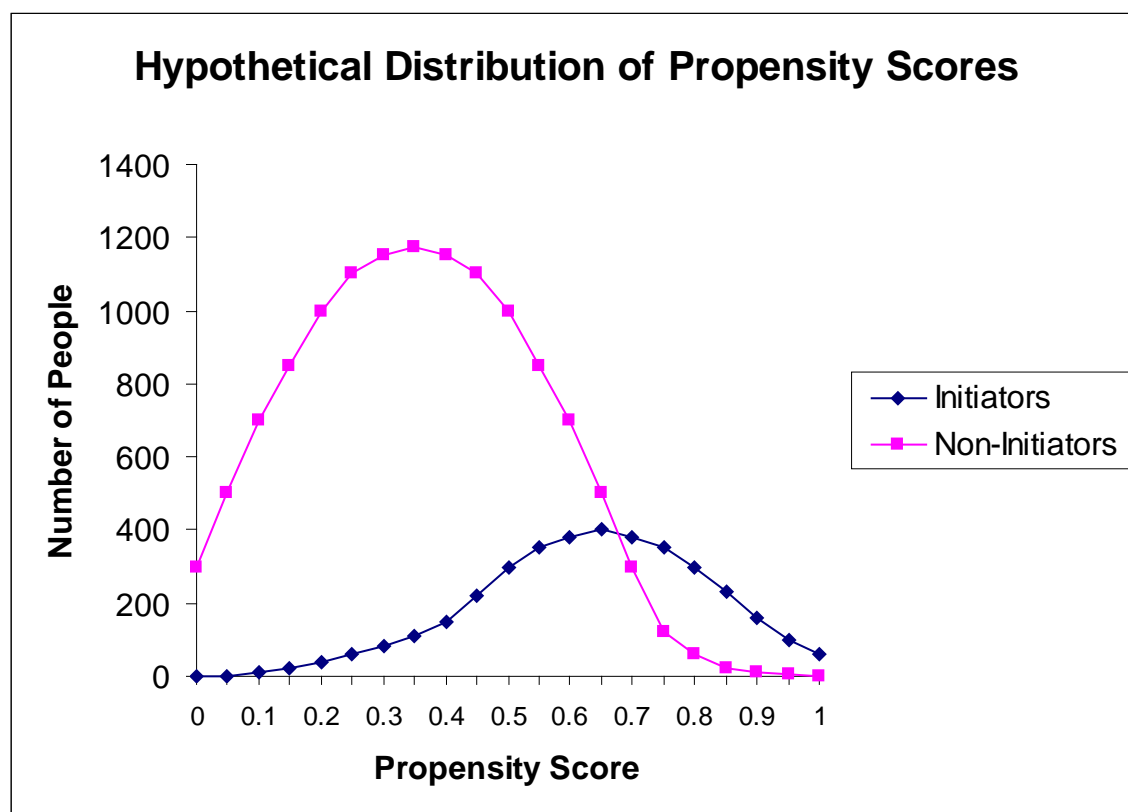
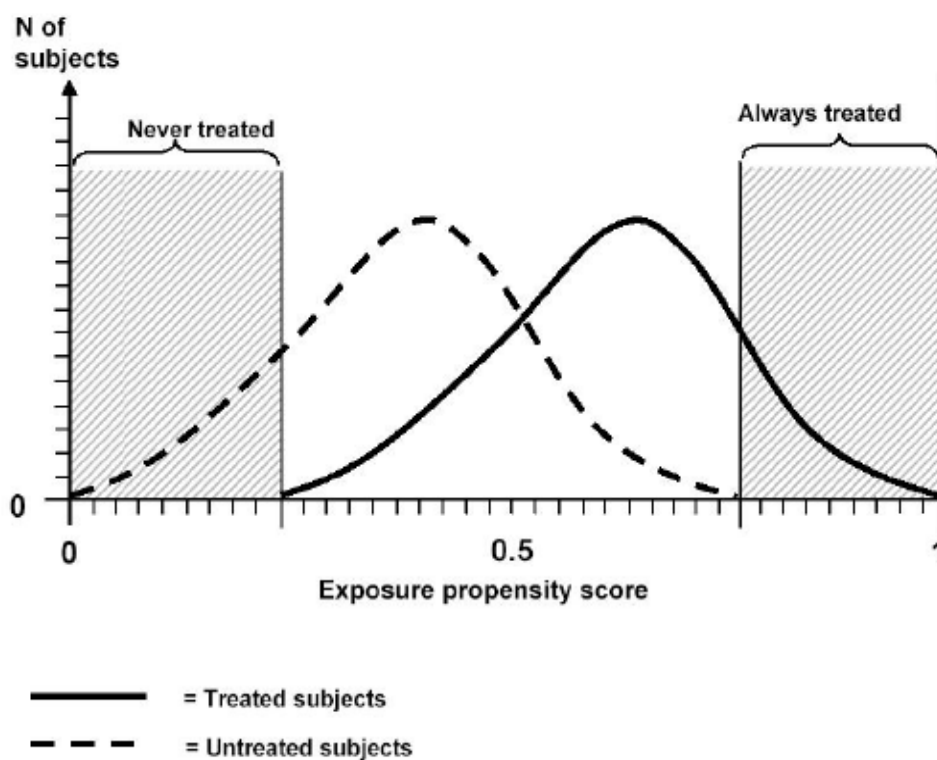
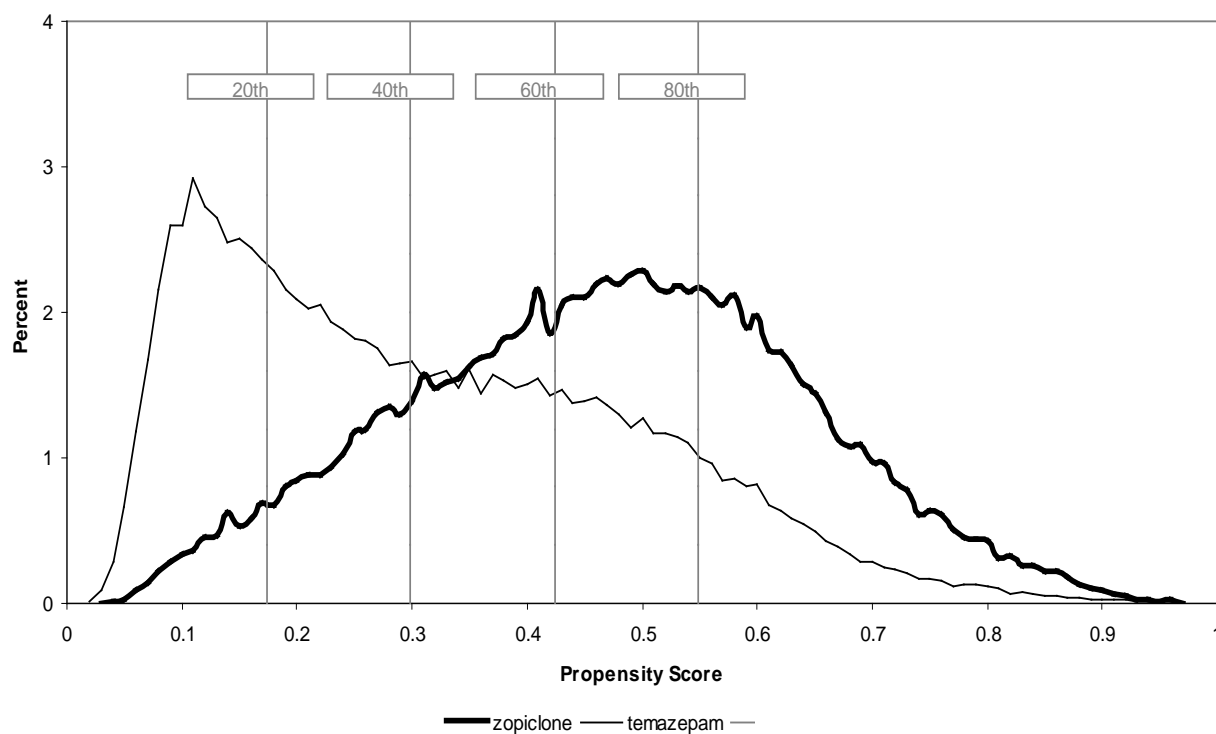


Figure 3. Illustration of propensity scores and non-overlap



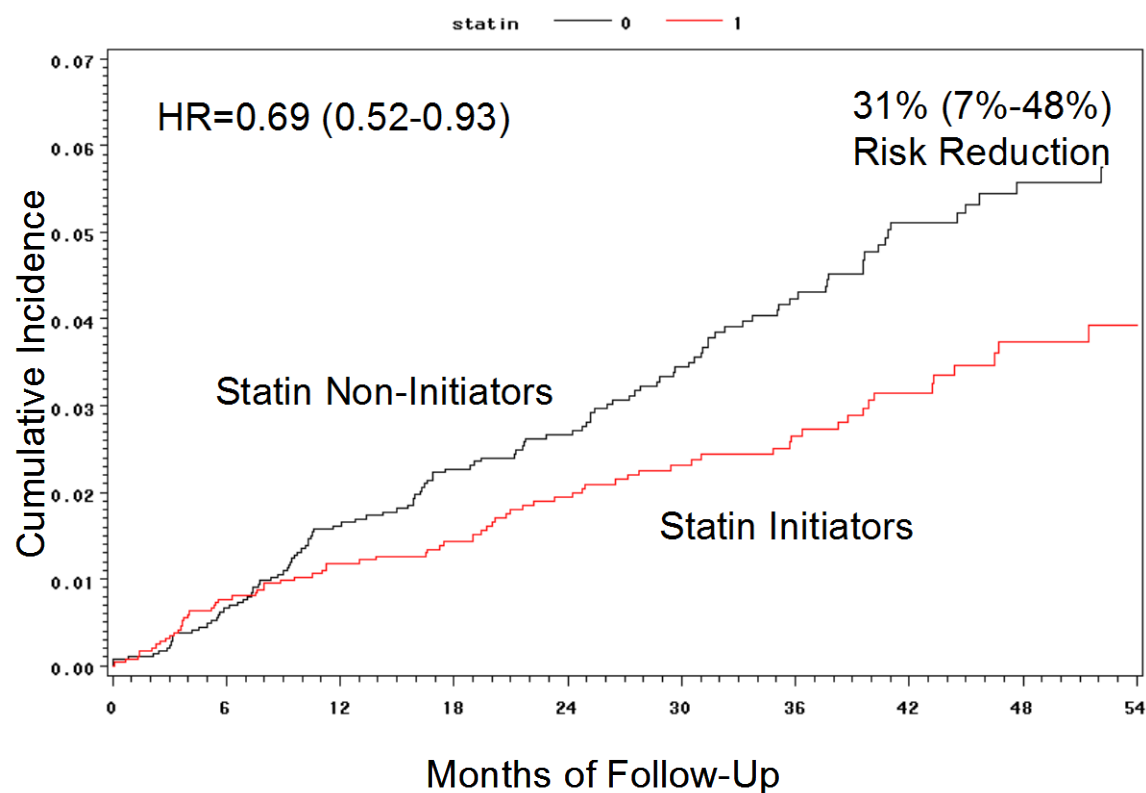
Source: Mosby, Inc., 2007. Used with permission. Schneeweiss S. Developments in post-marketing comparative effectiveness research. Clin Pharmacol Ther 2007 Aug;82(2):143-56.

Figure 4. Propensity score distribution illustrating stratification



Source: John Wiley & Sons, Inc., 2005. Used with permission. Eng PM, Ziyadeh N, Nordstrom BL, et al. Incidence of selected outcomes among matched cohorts of initiators of racemic zopiclone, temazepam, and zolpidem in the General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2005;14:S22-3.

Figure 5. Matched analysis presentation of results



Source: Reprinted from--Seeger JD, Walker AM, Williams PL, et al. A propensity score-matched cohort study of the effect of statin, mainly fluvastatin, on the occurrence of acute myocardial infarction. *Am J Cardiol* 2003;92:1447-51. Copyright 2003, with permission from Elsevier.

Appendix A. Bibliography

Propensity Score Methods Papers

Austin PC, Grootendorst P, Normand ST, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine* 2007;26:734-753.

Austin PC, Grootendorst P, Normand ST, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in Medicine* 2007;26:754-768.

Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine* 2007; 26: 3078-3094.

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *American Journal of Epidemiology* 2006;163:1149-1156.

Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology* 2003;158:280-287.

Eng PM, Seeger JD, Loughlin J, Clifford CR, Mentor S, Walker AM. Supplementary data collection with case-cohort analysis to address potential confounding in a cohort study of thromboembolism in oral contraceptive initiators matched on claims-based propensity scores. *Pharmacoepidemiol Drug Saf.* 2008;17:297-305.

Connors Jr AF, Speroff T, Dawson NV, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA* 1996;276:889-897.

D'Agostino Jr, RB. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 1998;17:2265-2281.

Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol* 2008;167:523-9.

Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology* 2006;98:253-259.

Harrell FE, Lee KL, Matchar DB, Reichart TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treatment Reports* 1985;69:1071-1077.

Imai K, van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *JASA* 2004;99:854-866.

Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000;87:706-10.

Joffe MM, Rosenbaum PR. Propensity scores. *American Journal of Epidemiology* 1999;150:327-33.

Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology* 2006;163:262-270.

Mansson R, Joffe MM, Sun W, Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology* 2007;166:332-339.

Parsons L. 2001. <http://www2.sas.com/proceedings/sugi26/p214-26.pdf>.

Robins JM, Mark SD, Newey WK. Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics* 1992;48:479-495.

Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association* 1987;82:387-94.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.

Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*. 1997;127:757-763.

Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine* 2007;26:20-36.

Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;14:680-686.

Schneeweiss S. Developments in Post-marketing Comparative Effectiveness Research. *Clin Pharmacol Ther*. 2007 Aug;82(2):143-56

Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology* 2005;58:323–337.

Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data. *Epidemiology* 2009;20:512–522.

Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety* 2008;17:546-555.

Sturmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology* 2005;162:279-289.

Sturmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration—a simulation study. *American Journal of Epidemiology* 2007;165:1110-1118.

Sturmer 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 specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology* 2006;59:437-447.

Propensity Score Clinical Papers

McAfee AT, Ming EE, Seeger JD, Quinn SG, Ng EW, Danielson JD, Cutone JA, Fox JC, Walker AM. The comparative safety of rosuvastatin: a retrospective matched cohort study in over 48,000 initiators of statin therapy. *Pharmacoepidemiol Drug Saf* 2006;15(7):444-53.

Seeger JD, Kurth T, Walker AM. Use of propensity score technique to account for exposure-related covariates. An example and lesson. *Medical Care* 2007;45(10):S143-8.

Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. *Pharmacoepidemiol Drug Safety* 2005;14(7):465-76.

Seeger JD, Walker AM, Williams PL, Saperia GM, Sacks FM. A Propensity Score–Matched Cohort Study of the Effect of Statin, Mainly Fluvastatin on the Occurrence of Acute Myocardial Infarction. *Am J Cardiol* 2003;92:1447-1451.

Eng PM, Ziyadeh N, Nordstrom BL, Caron J, Amato D, Seeger JD. Incidence of Selected Outcomes among Matched Cohorts of Initiators of Racemic Zopiclone, Temazepam, and Zolpidem in the General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2005;14:S22-3.

Cole JA, Zhu S, Russo LJ, Fife D, Walker AM. Use of Propensity Scores to Identify Possible Medication Dispensing Errors among Patients with Alzheimer's Disease. *Pharmacoepidemiol Drug Saf* 2005;14:S21.

Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, and Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of non-uniform effect. *Am J Epidemiol* 2006; 163(3):262-270.

Hurst FP, Bohen EM, Osgard EM, et al. Association of Oral Sodium Phosphate Purgative Use with Acute Kidney Injury. *Am J Soc Nephrol* 2007;18:3192-3198.

McKee S et al. Cocaine use is associated with an increased risk of stent thrombosis after percutaneous coronary intervention. *Am Heart J* 2007;154:159-64.

Ahmed A et al. Chronic kidney disease associated mortality in diastolic versus systolic heart failure: a propensity matched study. *Am J Cardiol*;99:393-398

McKee et al. Use of epidural anesthesia and the risk of acute postpartum urinary retention. *Am J Obstet Gynecol* 2007;196:471.e1-472.e5

H.D. Aronow et al., "In-Hospital Initiation of Lipid-Lowering Therapy after Coronary Intervention as a Predictor of Long-Term Utilization: A Propensity Analysis," *Arch Intern Med* 163, no. 21 (2003).

Disease Risk Score Methods Papers

Arbogast PG, Kaltenbach L, Ding H, Ray WA. Adjustment for multiple cardiovascular risk factors using a summary risk score. *Epidemiology* 2008;19:30-37.

Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Statistical Methods in Medical Research* 2009; 18: 67-80.

Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. *Journal of Clinical Epidemiology* 1989;42:317-324.

Hansen BB. The prognostic analogue of the propensity score. *Biometrika*. 2008; 95: 481-488.

Miettinen OS. Stratification by a multivariate confounder score. *American Journal of Epidemiology* 1976;104:609-620.

Pike MC, Anderson J, Day N. Some insights into Miettinen's multivariate confounder score approach to case-control study analysis. *Epidemiology and Community Health*. 1979;33:104-106.

Sturmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *American Journal of Epidemiology*. 2005;161:891-898.

Disease Risk Score Clinical Papers

Graham DJ, Campen D, Hui R, Spence M, Cheetham C, Levy G, Shoor S, Ray WA. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet* 2005;365:475–481.

Ray WA, Meredith S, Thapa PB, Meador KG, Hall K, Murray KT. Antipsychotics and the risk of sudden cardiac death. *Archives of General Psychiatry* 2001;58:1161–1167.

Ray WA, Stein CM, Hall K, Daugherty JR, Griffin MR. Non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease: an observational cohort study. *Lancet* 2002;359:118–123.

Ray WA, Stein CM, Daugherty JR, Hall K, Arbogast PG, Griffin MR. COX-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. *Lancet* 2002;360:1071–1073.

Ray WA, Meredith S, Thapa PB, Hall K, Murray KT. Cyclic antidepressants and the risk of sudden cardiac death. *Clinical Pharmacology & Therapeutics* 2004;75:234–241.

Ray WA, Murray KT, Meredith S, Narasimhulu SS, Hall K, Stein CM. Oral erythromycin and the risk of sudden death from cardiac causes. *N Engl J Med* 2004;351:1089–1096.

Ray WA, Chung CP, Stein CM, Smalley WE, Hall K, Arbogast PG, Griffin MR. Risk of peptic ulcer hospitalizations in users of NSAIDs with gastroprotective cotherapy versus coxibs. *Gastroenterology* 2007;133:790–798.

Miscellaneous Papers

Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine* 1989; 8: 771–783.

Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968;24:295–313.

Graham DJ, Staffa JA, Shatin D, Andrade SE, Schech SD, La Grenade L, Gurwitz JH, Chan KA, Goodman MJ, Platt R. Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. *JAMA* 2004;292:2585–90.

Expert Panel on the Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA*. 2001;285:2486–2497.

Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf* 2006; 15(5): 291–303.