USING HIGH-DIMENSIONAL DISEASE RISK SCORES IN COMPARATIVE EFFECTIVENESS RESEARCH OF NEW TREATMENTS

Richard Wyss

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Epidemiology.

Chapel Hill
2015

Approved by:

Til Stürmer

M. Alan Brookhart

Michele Jonsson Funk

Cynthia J. Girman

Ross J. Simpson

ABSTRACT

RICHARD WYSS: Using High Dimensional Disease Risk Scores in Comparative Effectiveness Research of New Treatments
(Under the direction of Dr. Til Stürmer)

Nonexperimental research using automated healthcare databases can supplement randomized trials to provide both clinicians and patients with timely information to optimize treatment decisions. These studies, however, are susceptible to confounding and require design and statistical methods to control for large numbers of confounding variables. The propensity score (PS), defined as the conditional probability of treatment given a set of covariates, has become increasingly popular for controlling large numbers of covariates in pharmacoepidemiologic studies. During early periods after the introduction of a new treatment, however, accurately modeling the PS can be difficult because of rapid change over time in drug prescribing patterns and few exposed individuals. A historically estimated disease risk score (DRS), which summarizes covariate associations with the outcome absent of exposure, has been proposed as an alternative to PSs for controlling large numbers of covariates during these periods. Little is known about the performance and potential benefits of using DRSs for confounding control when evaluating the comparative effectiveness of newly marketed drugs.

In this study, we examined the benefits and challenges of using historically estimated DRSs compared to PSs when controlling for large numbers of covariates during early periods of

drug approval. We further evaluated novel strategies for determining the validity of fitted DRS models in their ability to control confounding. We investigated these methodological questions using Monte Carlo simulations and empirical data. The empirical analyses included 20% and 1% samples of Medicare claims data to compare the new oral anticoagulant dabigatran with warfarin in reducing the risk of combined ischemic stroke and all-cause mortality in older populations.

When PS distributions are separated, DRS matching can improve the precision of effect estimates and allow researchers to evaluate the treatment effect in a larger proportion of the treated population. However, accurately modeling the DRS can be challenging compared to the PS. When evaluating the validity of DRS models, measures of predictive performance do not always correspond well with reduced bias in treatment effect estimates. Calculating the pseudo bias within a “dry run” analysis can provide a more direct measure for assessing the ability of fitted DRS models to control confounding.

To my wife, Annah, for your love and support

and

To my parents, Rick and Arlene, for your support and encouragement

ACKNOWLEDGEMENTS

I would first like to thank my adviser, Til Stürmer, for his guidance and mentoring. His example and mentorship has been essential in helping me grow as a student of pharmacoepidemiology. Thank you for facilitating many valuable research opportunities and for always being willing to make time in your schedule to answer questions and discuss methodological concepts. I am very grateful for the opportunity that I have had to work with you these past years.

I would like to thank the members of my dissertation committee: Alan Brookhart, Michele Jonsson Funk, Cynthia J Girman, and Ross Simpson. Thank you for your encouragement and the many hours you invested in providing your guidance and expertise.

I would also like to thank Alan Ellis for the many hours that he has invested in providing his expertise in the projects that we have worked on together and Virginia Pate for her help and expertise in programming. I am also very grateful for the friendships and support from my classmates in the pharmacoepidemiology program.

Lastly, I would like to express my deep appreciation and love for my wife Annah and son Daniel for their love, patience and support. I would also like to express my deep gratitude to my parents for their love and support during my time at UNC and beyond.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AF	Atrial fibrillation
CVD	Cardiovascular disease
DRS	Disease risk score
ICD-9	International classification of disease, 9th edition
HR	Hazard ratio
PS	Propensity score
RCT	Randomized controlled trial
SMR	Standardized mortality ratio

CHAPTER 1

BACKGROUND AND SPECIFIC AIMS

1.1 Specific Aims

Controlling large numbers of confounding variables presents unique challenges when evaluating new treatments in comparative effectiveness research. Summary scores, which reduce covariate information to a single dimension, have become increasingly popular for controlling large numbers of baseline covariates. The propensity score, defined as the conditional probability of treatment given a set of measured covariates, has become the most widely used summary score in pharmacoepidemiologic research.[1, 2] It has been hypothesized, however, that modeling the PS during early periods of treatment introduction can be difficult as the number of individuals receiving the new treatment can be small and factors affecting treatment assignment can change rapidly during early periods of dissemination.[3, 4]

An alternative summary score to the PS is the prognostic score, also known as the disease risk score (DRS). Instead of modeling covariate associations with treatment, the DRS models the relationship between covariates and the potential outcome under the control or comparator treatment.[5-7] Because factors affecting disease risk are more stable over time, it has been proposed that modeling the DRS within historical data prior to treatment introduction can provide an alternative to the PS for controlling large numbers of covariates when evaluating new treatments.[3] However, DRSs have not been widely used and the

validity and potential benefits of a historically estimated DRS remain unclear. In this study we used both Monte Carlo simulations and empirical data to examine the benefits of using a historically estimated DRS for controlling large numbers of covariates when evaluating new treatments. We further develop and evaluate methods for assessing the validity of DRS models directly in their ability to control for confounding. Empirical analyses focused on comparing of the new oral anti-coagulant dabigatran with warfarin in preventing combined ischemic stroke and all-cause mortality using Medicare claims data. There has been recent interest in the use of new oral anticoagulant drugs as an alternative to warfarin for patients with atrial fibrillation.[8-10] Although clinical studies evaluating these new drug classes have shown promising results,[10] their comparative effectiveness in real world patient populations remain largely unknown. With limited data at the beginning of drug approval, estimation of the PS as a function of large numbers of covariates can be problematic. In contrast, estimating the DRS within historical data could allow researchers to effectively control for large numbers of covariates immediately after drug approval, when data on safety is most important. Improved methods for confounding control during early periods of drug approval and evolving drug therapies can enhance treatment decisions for healthcare providers and the patient community.

Aim 1: Use both Monte-Carlo simulations and empirical analyses to better understand potential benefits of using a historically estimated DRS when controlling large numbers of confounding variables during early periods of drug approval. The empirical analysis will focus on evaluating the comparative effectiveness of the new oral

anticoagulant medication dabigatran compared with warfarin in preventing ischemic stroke and all-cause mortality in patients with atrial fibrillation in the Medicare population.

Rationale: There remains little evidence and understanding of the benefits of a historically estimated DRS compared with traditional PS methods in practice. It remains unclear what specific advantages the DRS provides over traditional PS methods.

Aim 2: Use both simulations and substantive analyses to develop and compare novel strategies for evaluating risk models in their ability to control for confounding.

Rationale: Accurately modeling the DRS presents unique challenges that are not shared by traditional outcome regression modeling or PS estimation. These difficulties highlight the importance of evaluating the validity of the fitted risk model in its ability to control for confounding. The validity of fitted DRS models has primarily been assessed through measures of predictive performance which give an indirect assessment of the ability of the DRS to control for confounding. Recent methods that use the control population to create pseudo treatment and pseudo control groups have been proposed as a more direct measure for assessing the validity of risk models in terms of confounding control. It remains unclear what metrics are optimal for evaluating risk models.

1.2 Comparative Effectiveness Research

Randomized controlled trials (RCTs) are the gold standard for evaluating the performance of a treatment or drug.[11] The strict design of RCTs, including randomization and blinding, ensures the internal validity and effective control of variables that may bias results. However, information provided by RCTs can be limited for real world clinical practice. RCTs can fail to detect rare outcomes and long-term effects due to smaller sample

sizes and shorter durations of treatment.[12-15] RCTs can also have limited generalizability due to restrictions placed on study participation including age, comorbidity and co-medication.[13, 14] Further, RCTs often assess efficacy versus placebo rather than versus an alternative treatment for the same indication.

Comparative effectiveness research has gained considerable attention in recent years. It is becoming increasingly recognized that RCTs cannot address every question regarding treatment decisions for patients in real world clinical practice. Large automated healthcare databases, such as administrative data and electronic medical records, are increasingly being used to evaluate drug performance and safety.[15] Compared with RCTs, observational studies are better suited to provide information on drug utilization as well as benefits and harms of drugs in real world settings with populations covering a wide range of patient characteristics.[11, 14, 15] Automated databases can provide valuable information on the real time performance of medical treatments. This is critical for the active surveillance of drug effects in real world populations.[16] Observational studies using healthcare and administrative datasets complement RCTs to improve healthcare providers' decisions regarding drug and treatment choices.[15] However, the evaluation of drug effects using observational data is susceptible to both measured and unmeasured confounding that is caused by the lack of randomization. The validity of studies using automated databases is limited by the ability of current statistical and epidemiologic methods to effectively control for large numbers of confounding variables.

1.2.1 Current Methodological Limitations

The continual development and application of novel methods is essential to reduce bias in observational studies and more accurately address important public health issues regarding drug effects and treatment decisions in real world populations. The development of improved methods for confounding control is particularly needed during early periods of drug approval.[3] During these early periods, there is often rapid change over time in drug prescribing patterns or in the use of a treatment.[4] New users of a recently approved drug will often have different patient characteristics than new users of the same drug after the drug has been on the market for an extended period of time. Such changes over early periods of drug approval present significant challenges for comparative effectiveness research and can make rapid response for drug safety difficult. Developing improved methods to control for confounding during early periods of drug approval is needed to provide the best evidence for treatment decisions during these early stages.

Standard methods for confounding control have traditionally consisted of multiple regression models. Although useful in many situations, these methods are limited for studies involving large numbers of confounders due to computational complexity, the high likelihood of model misspecification and the limited ability to model more complicated functional relationships such as interactions and higher order terms for rare outcomes.[7] To address these limitations, methods that collapse the information of a large number of covariates into a single-dimensional summary score and then use this summary measure for confounding control have become increasingly popular.[1, 6, 17]

1.3 Summary Scores and Causal Inference

1.3.1 Counterfactual Framework

The Neyman-Rubin counterfactual framework provides a formal framework for researchers to determine causal effects in both experimental and non-experimental studies.[18-20] Under the counterfactual model for causal inference, each person in the study population has a potential outcome corresponding to each possible treatment level. For example, if T represents a dichotomous treatment, then Y_1 represents the potential response had the individual received treatment and Y_0 the potential response had the individual received the control or comparator treatment. In practice, only one of the potential outcomes is observed for each individual. The observed response, Y, has the following relationship with the potential outcomes

$$Y = (T)Y_1 + (1 - T)Y_0.$$

The treatment variable, T, is said to have a causal effect on the observed outcome, Y, for a given individual if $Y_1 \neq Y_0$. For a population of individuals, T has an average causal effect on Y within the entire population if $E[Y_1] \neq E[Y_0]$, where $E[Y_1]$ and $E[Y_0]$ represent the expected or average value of the random variables Y_1 and Y_0 respectively.

A fundamental obstacle in non-experimental studies is estimating treatment effects in the presence of confounding factors. If controlling for a set of baseline covariates, X, results in treatment assignment being independent of potential outcomes, then average treatment effects in the population are identifiable.[19] Known as the strongly ignorable treatment assignment assumption, this condition is formally expressed as $(Y_1, Y_0) \perp T|X$ where $\perp$ denotes independence of random variables and $|$ denotes conditional on.

Heckman[21] showed that when estimating the treatment effect on the treated, the strongly ignorable treatment assignment assumption is unnecessarily restrictive and the weaker condition, $Y_0 \perp T|X$, is sufficient. Known as the weakly ignorable treatment assignment assumption, this condition is sufficient to identify average treatment effects on the treated population.[5, 21]

1.3.2 Propensity Score Methods

For a dichotomous treatment, T, the propensity score is defined as the conditional probability of treatment assignment given a set of baseline covariates, X. Formally expressed as $\varphi(X) = E[T|X]$, Rosenbaum & Rubin[1] show that conditioning on the PS results in covariates being independent of treatment assignment, formally denoted as $X \perp T|\varphi(X)$. If treatment assignment is strongly ignorable given a set of baseline covariates, i.e. $(Y_1, Y_0) \perp T|X$, Rosenbaum & Rubin[1] show that treatment assignment is strongly ignorable given the PS, i.e. $(Y_1, Y_0) \perp T|\varphi(X)$. This independence allows for the identification of average treatment effects in the full study population or average treatment effects in subgroups of the study population, e.g., the treated population. If treatment assignment is weakly ignorable, i.e. $Y_0 \perp T|X$, then conditioning on the propensity score satisfies $Y_0 \perp T|\varphi(X)$ allowing for the identification of average treatment effects in the treated population.

The development and advancement of PSs in various applications and settings has been a key factor for improved methodological standards and validity when evaluating drug effects in non-experimental settings. However, the performance of PSs is limited in certain settings. It has been hypothesized that PSs may not perform well for studies involving rare

or emerging treatments.[22] Factors affecting treatment assignment are not necessarily biological in nature and are more likely to vary and change over time compared to factors that affect the outcome or disease.[3, 6, 7] For example, physicians becoming more familiar with a new treatment may extend the indication to patients with less severe disease or more severe co-morbidities. Other situations where transient factors affect indication for treatment include a newly approved treatment quickly diffusing through the market and the issuance of black box warnings. Modeling the PS can potentially be difficult when factors affecting treatment assignment change over short periods of time.

The PS is also not a natural measure to evaluate treatment effect heterogeneity. When making treatment decisions, clinicians are almost always concerned about how the effect of a treatment varies over various patient profiles affecting the risk for the outcome of interest (e.g., 10 year risk for cardiovascular disease based on the Framingham risk score). Although the PS allows us to detect and account for treatment effect heterogeneity, it does not provide the best information for health care providers in determining what subgroups of the patient population are most likely to benefit from a given treatment regime.

1.3.3 Disease Risk Score Methods

The disease risk score (DRS) has been shown to be a valid alternative to PSs for controlling large sets of confounders.[5, 7, 23, 24] Originally introduced by Peters in 1941[25] as a way to reduce dimensionality when matching, the DRS has been used by a variety of researchers to control for confounding and assess treatment effect heterogeneity.[26, 27] The DRS is similar to the PS in that the DRS summarizes the information of a large number of

variables with a single-dimensional score.[27] Unlike PSs, however, DRSs summarize the associations of baseline covariates with the potential outcome under the control therapy instead of treatment, i.e., the risk for the outcome.

Despite the early introduction of DRSs, their use was inhibited in part due to an early study by Pike, et al.[28] that examined their statistical properties.[6, 28] Pike demonstrated that adjustment for the DRS can result in exaggerated statistical significance of effect estimates.[28] After reexamining these findings, Cook & Goldman[29] found that this exaggeration is small except when there is a very strong correlation between confounders and the exposure (correlation coefficient exceeding 90%), which is unlikely to occur in practical settings.[6, 7, 29] Leacy further explains that this exaggeration in statistical significance is due to issues of model misspecification rather than the statistical properties of the DRS.[30]

Recently, Hansen[5] has solidified the theoretical foundation for the use of DRSs in causal inference. Hansen showed that the DRS acts as a prognostic balancing score that can yield valid effect estimates with a causal interpretation.[5] Formally Hansen defines the prognostic score, or disease risk score, as any scalar or multi-dimensional function of X that satisfies the condition $Y_0 \perp X|\psi(X)$.[5] In other words, conditioning on the DRS results in a form of covariate balance where the potential response under control is independent of a set of measured covariates, X. If the outcome follows a generalized linear model, Hansen[5] shows that one possible prognostic score, or DRS, is the linear predictor of Y_0, or the conditional mean of Y_0 given X (i.e., $E[Y_0|X]$). Hansen[5] further shows that if treatment assignment is weakly ignorable given a set of baseline covariates, i.e. $Y_0 \perp T|X$, then conditioning on the DRS is sufficient to satisfy $Y_0 \perp T|\psi(X)$ allowing for unconfounded

estimates of the treatment effect in the treated population through stratification or matching on the DRS.

Using simulations, Arbogast and Ray[7] evaluated the properties of effect estimates when applying DRSs. Their study showed the DRS to perform similar to PSs and outcome regression models for the settings evaluated.[7] Stürmer et al.[24] and Cadarette et al.[31] used data from Medicare recipients to evaluate the performance of disease risk scores compared to PS methods and traditional outcome regression in real world settings. In these examples, results were similar from the application of DRSs, PSs, or traditional multivariable regression.

Due in part to this recent theoretical work and evaluation of the properties of DRSs using both simulations and empirical data, there has been increased interest in the application of DRSs to evaluate drug performance.[23] Although generally not superior to PSs, the DRS can be advantageous to PSs for controlling confounding in certain settings.[24] For example, studies with rare exposures (e.g. emerging therapies) and studies involving multiple therapies can benefit from DRSs which model covariate associations with the outcome rather than treatment.[7, 31] Further, DRSs provide a natural measure to evaluate treatment effect heterogeneity. Evaluating treatment effect heterogeneity across the distribution of disease risk provides a straightforward approach for clinicians to identify subgroups of patients that are most likely to benefit from the treatment, thereby improving treatment decisions made by healthcare providers.[3]

Despite this recent attention, there remain many unanswered questions regarding the use of DRSs in practice. A fundamental challenge in applying DRSs is understanding how

these summary measures should be estimated. Although various strategies for estimation have been proposed, there remains uncertainty regarding which estimation strategy is optimal in diverse settings. This uncertainty is particularly acute for studies using large administrative datasets to evaluate the comparative effectiveness of drugs because there have been relatively few applications of disease risk scores in these settings. Multiple researchers have expressed the need for further empirical and simulation studies to clarify the application of DRSs in real world practice.[6, 31]

Traditionally, the DRS has been estimated in two ways. The first is to fit a regression model to untreated individuals within the cohort and then use this model to predict the disease risk for all individuals within the full cohort. The second is to fit a regression model to the full cohort (both treated and untreated) as a function of baseline covariates and treatment, and then estimate the disease risk for each individual after setting treatment status to untreated.[3, 5, 7]

Fitting the DRS to the full cohort benefits from increased sample size, but requires accurately modeling the relationship between the treatment and outcome.[5] Hansen shows that when estimating the DRS within the full cohort, incorrectly modeling the modification of treatment by baseline covariates (i.e. disease risk) can result in estimated scores that carry information about the true treatment effect. This non-ancillarity in the estimated scores can obscure the effect estimate when used for stratified or matching analysis. Correctly modeling treatment effect heterogeneity by disease risk can be difficult, particularly for large numbers of covariates. Therefore, Hansen recommends using only the untreated cohort for estimation of the disease risk.[5] However, accurate estimation of the

DRS using only the untreated cohort presents its own challenges. Fitting the DRS only within the untreated cohort can introduce bias by substantially increasing the potential for overfitting the model.[3, 5]

Recently, alternative strategies for estimating DRSs that use data from outside the defined new user cohort have been proposed. Both Hansen and Glynn discuss potential advantages of using outside data to estimate the DRS.[3, 5] Hansen explains that estimating the DRS within an alternate sample of controls can avoid the complications of overfitting that can occur when using same-sample estimation. Glynn suggests that out-of-sample estimation of the DRS can be particularly advantageous when evaluating evolving drug therapies because the first patient receiving the new treatment can be matched to a concurrent patient receiving the old treatment based on the estimated disease risk.[3]

In contrast with predictors of treatment, factors that predict outcome or disease occurrence are more likely to be biological in nature and less likely to vary over time and be impacted by physician decisions which can be difficult to identify.[3, 6] Because the DRS is likely to be stable over time and across populations, Glynn proposes that disease risk can be accurately estimated from either a separate population, or the same population but with historical data from a period prior to the current study period.[3]

During early periods of drug approval, there is usually limited data to accurately estimate either the PS or DRS as a function of large numbers of covariates, particularly for studies involving rare outcomes or rare exposure. Further, during early periods of a newly introduced drug, a well-defined PS may not exist due to evolving factors affecting treatment

assignment. Out-of-sample estimation for the DRS potentially avoids these challenges by using a sample with sufficient data to accurately estimate the disease risk.

While having the potential to improve confounding control for the early evaluation of treatment therapies, the performance and potential benefits of historical estimation of disease risk is not well established. Using observations, or information, from historical data can present important challenges. Covariate assessments and coding practices can change over time making a historically estimated DRS not generalizable to future time periods and populations.

The challenges outlined above when estimating the DRS highlight the importance of evaluating the validity of fitted DRS models in their ability to control for confounding. If prognostic balance could be evaluated within the full study population, then measures of prognostic balance could be used to evaluate the validity of fitted DRS models in a similar way that measures of covariate balance across treatment groups are used to evaluate PS models. Prognostic balance, however, can only be evaluated within individuals receiving the comparator treatment where the potential outcome under control, Y_0, is observed. It is unclear how well measures of prognostic balance within only the control group correspond to a reduction in bias in the estimated treatment effect. Measuring prognostic balance only within the comparator group can potentially reward models that are overfit to the control population. Further, in the presence of unmeasured confounding, the DRS does not result in prognostic balance within subgroups of treatment, but only marginally within the entire population.[5, 32] Measures of prognostic balance only within the control population can potentially lead to an incorrect assessment of the specified DRS model.

The inability to evaluate prognostic balance within the full study population has led to researchers evaluating DRS models primarily through measures of predictive performance, such as the c-statistic and goodness of fit tests. However, it is unclear how well measures of predictive performance correspond with the ability of DRS models to control confounding. When fitting PS models, previous studies have shown that measures of predictive performance do not always correspond well with reduced bias in the estimated treatment effect.[33, 34] Little attention has been given to determining what metrics are optimal for evaluating fitted DRS models.

1.4 Anti-coagulant medications and cardiovascular disease in patients with atrial fibrillation

Atrial fibrillation (AF) is the most common cardiac dysrhythmia in the United States and is a growing public health concern.[35, 36] AF is an established risk factor for stroke, cognitive dysfunction, and premature death.[37-39] It has been estimated that AF accounts for up to 15% of strokes in people of all ages and up to 30% of strokes in people over 80 years of age.[40, 41] Results from the Framingham Heart Study showed the prevalence of AF to be 6% and estimated that the lifetime risk for developing AF is approximately 25% for both men and women 40 years of age and older.[39] The Framingham Heart Study has further shown that the risk for AF increases with age. As the elderly population increases in the United States, the prevalence of AF is expected to increase substantially over the next few decades.[42]

Standard practice for treating patients with AF includes treatment with an anti-coagulant for the prevention of thromboembolic events.[8] Long term oral anticoagulant treatment options for patients with AF have only included long-acting VKAs (warfarin).[8]

Warfarin has been the most widely used VKA and has been shown to decrease rates of stroke in patients with AF in several trials.[8, 43] Warfarin reduces blood coagulation by inhibiting vitamin K-dependent clotting factors. However, the magnitude of the effect on these factors is variable and difficult to predict. Warfarin is a drug with a very narrow therapeutic window and pronounced inter- and intraindividual variation of effects on coagulation. Warfarin thus needs intense monitoring of its effects on coagulation based on the INR that needs to be kept within a narrow range. Both ineffectiveness (too low INR) and increased risk of bleeding (too high INR) are quite common. Consequently, AF patients being treated with warfarin need to be closely monitored to assure that patients are attaining an effective dose range. Several studies have shown that as many as 45% of patients on warfarin are not within the therapeutic range for a sufficient period of time.[8, 10, 44] Although most patients have been shown to benefit from warfarin, concern for these potential complications and adverse bleeding events has often led to an underuse of anticoagulant medications among persons with AF.[45, 46] It is estimated that warfarin has only been used by 30-60% of appropriate patients with AF.[45, 47, 48]

New oral anti-coagulant medications that focus on inhibiting a specific factor in the coagulation pathway have recently been developed to overcome the shortcomings of warfarin.[43] There are numerous oral anticoagulant agents in development. The most advanced in clinical research belong to two drug classes: direct thrombin inhibitors (DTIs) and factor Xa (FXa) inhibitors.[8] Clinical studies evaluating these drug classes in patients with AF have shown promising results. Dabigatran, a direct thrombin inhibitor, showed reduced rates of stroke or systemic embolism in select patients with AF when compared to warfarin

in the RE-LY trial.[49] The ROCKET-AF trial showed the oral FXa inhibitor Rivaroxaban to be noninferior to warfarin in reducing rates of stroke and produced no significant difference in the risk of major bleeding.[41] In the ARISTOTLE randomized trial comparing the direct FXa inhibitor apixaban to warfarin in patients with AF, apixaban was also shown to be noninferior to warfarin in reducing stroke while resulting in fewer major bleeding events.[50] Unlike warfarin, clinical data have further shown that these novel agents have predictable pharmacokinetic, pharmacodynamics, and anticoagulant response thereby. The predictable performance of these newer agents reduces the need for dose adjustments and frequent monitoring of coagulation parameters.[8, 9, 51] The abundant clinical data supporting the efficacy of these novel oral anticoagulant agents has led to the recent FDA approval of dabigatran in October 2010, rivaroxaban in November 2011, and apixaban in December 2012. In addition to these oral anticoagulants, there are several additional novel anticoagulant agents in advanced development.

As novel anticoagulants become more widely used, it is possible that there will be a paradigm shift in the prescribing of anticoagulation treatments for patients with AF.[8] Although clinical data suggest increased potential for achieving an effective dose range with these newer agents without the need for frequent monitoring, the implications on population level practice have not been adequately evaluated. Ansel[52] discusses limitations of these clinical trials highlighting important inclusion/exclusion criteria that may not generalize to these new oral anticoagulants having the same performance in real world AF patients. A systematic review comparing new oral anticoagulants (dabigatran, rivaroxaban and apixaban) to warfarin concluded that these new agents had a lower risk for fatal

bleeding and hemorrhagic stroke, but an increased risk for gastrointestinal bleeding, a more common event in the elderly.[10] Differential performance of these new anticoagulants between young and elderly populations is not well established.

Many clinicians acknowledge that there are gaps in the current understanding of how these new medications perform in clinical practice.[52-54] Ansell[52] asserts that there are enough unknowns regarding the effects of these new oral anticoagulants that health care providers and patients should be cautious when using these medications as first line treatment. Observational research has the potential to supplement information provided by the randomized trials and improve our understanding of these new anticoagulants in real world practice and diverse patient populations. In a systematic review conducted by Adam et al,[10] the authors found that the observational literature on adverse events of new oral anticoagulants is sparse, consisting only of case reports. We seek to improve the information regarding the performance of recently introduced novel oral anticoagulant medications by using the previously described methods to thoroughly investigate their performance in elderly populations using Medicare data.

1.4.1 Early Evaluation of New Oral Anticoagulant Drugs for Patients with Atrial Fibrillation

Due to the very recent approval of these new oral anticoagulants, their evaluation in non-experimental settings is difficult, in part due to the limited data available to control for large sets of confounders. Out-of-sample estimation methods for disease risk are advantageous because these methods can potentially allow for the control of a large number of risk factors at the start of drug introduction. While Medicare data do not capture some important clinical variables, including blood pressure, they offer major advantages for

this kind of research based on the population-based nature (real world) and the ability to evaluate clinically relevant outcomes rather than intermediates within a very short time (because of the overall size of the population).[55, 56] This will potentially allow researchers to effectively evaluate the comparative effectiveness of these newer medications at earlier periods than previous methods.

CHAPTER 2

METHODS

2.1 Study Population

This study consisted of older individuals ages 65 years and older who were beneficiaries of the Medicare system. Medicare is a national insurance program for all Americans over the age of 65. Because all elderly Americans are entitled to these benefits, individuals receiving Medicare are likely to be representative of the general health care utilization of elderly adults in the US population. However, the Medicare data made available for research is limited to parts A, B, and D. Medicare part C, which includes Medicare Advantage plans, is administered by private insurance companies and is not made publicly available. Medicare part C covers approximately 25% of all Medicare beneficiaries and contains individuals who tend to have higher socioeconomic status. Furthermore, the study populations for the research questions in this study excluded individuals who did not participate in Medicare part D. The exclusion of Medicare part C recipients and individuals who do not participate in part D may affect the study population to be more representative of Medicare beneficiaries of lower socioeconomic status.

The Medicare data from the Center for the Medicare and Medicaid Services (CMS) Chronic Condition Data Warehouse from 2007 to 2012 are available at UNC. The CCW files include annual enrollment summary, inpatient, outpatient, skilled nursing facility, carrier (physician office visit), hospice, home healthcare agencies, durable medical equipment files

and prescription Part D event files. The prescription Part D files include the national drug code of the medication, service data, strength of the medication, days of supply, quantity dispensed, encrypted and unique prescriber identifier, unique and encrypted pharmacy identifier, generic drug name, and the benefit phase of the Part D event. All CCW files are linked by an encrypted and unique CCW identifier number for each beneficiary. We used 100% of patients nationwide who meet the inclusion and exclusion criteria.

2.2 Study Design

We included individuals who were continuously enrolled in the Medicare data at least 12 months before and through the end of the study period. A new user cohort study design was used to evaluate the described research questions.[11] We identified Rx claims for dabigatran and warfarin within Medicare A, B, and D claims data. We determined periods of new use after a pre-specified washout period. New users of dabigatran who had a prescription claim for the comparator drug during the specified washout period were excluded. We identified new users of dabigatran and warfarin between the years 2010-2012. The start of this time period corresponds to FDA approval for dabigatran.

We applied a new user cohort design when evaluating the previously described research questions to mitigate both measured and unmeasured confounding caused by indication for treatment and healthy users. Confounding caused by indication for treatment and healthy-users are two primary sources of bias in comparative effectiveness research. New user designs reduce the potential for healthy user bias by excluding prevalent users.[57] Prevalent users of a drug at baseline of follow-up are more likely to have systematic differences in the distributions of unmeasured risk factors for the outcome compared to

new initiators of the drug.[11, 57] In each of the analyses, we further reduced the magnitude of confounding caused by indication for treatment by comparing the defined exposure to an active comparator with similar indication.[58]

All decisions on exclusion criteria for cohort participation were made prior to treatment initiation and follow-up. Measured confounders were controlled at baseline using propensity score matching and disease risk score matching.

2.3 Choice of Outcomes

We considered a combined outcome of ischemic stroke and all-cause mortality. Ischemic stroke was defined as hospitalization with diagnostic codes in the principal or secondary positions (Primary Dx 430-434). Birman-Deych, et al[59] demonstrated that ICD-9 codes for coronary artery disease, stroke, heart failure, and hypertension had high specificity (>0.95), low sensitivity (<0.76), and a positive predicted value of 0.95 within the Medicare Part A data. They further demonstrate that miscoding of ischemic stroke events as hemorrhagic events was rare. Similar findings have been found by other studies evaluating the validity of ICD-9 codes in other large administrative databases.[60]

2.4 Variable Selection

For the empirical analysis, we selected a high-dimensional set of covariates using an algorithm that is similar in concept to the high-dimensional PS.[61] We first selected a reduced set of covariates a priori using expert knowledge. We then included an additional 200 empirically selected covariates that were identified within Medicare files containing medication claims, inpatient and outpatient diagnostic codes and procedural codes. When selecting the 200 additional covariates, we first identified the top 200 most prevalent codes

within each data dimension of the Medicare data (codes with a prevalence greater than 0.5 were subtracted from 1). Among the codes selected, we then identified the top 200 codes based on the strength of their univariate association (odds ratio) with the outcome.

2.5 Monte Carlo Simulations

We used Monte Carlo simulations to better understand and evaluate potential benefits of DRS matching vs PS matching for Aim 1. Simulations were also used in Aim 2 to compare various measures for evaluating the validity of the specified DRS models.

2.5.1 Simulations for Aim 1

The simulated causal structure for Aim 1 was motivated by the empirical example comparing new-users of dabigatran with warfarin in preventing ischemic stroke and all-cause mortality. We simulated 100 baseline covariates to reflect settings involving high-dimensional sets of covariates. Baseline covariates included a mixture of both continuous and dichotomous random variables. We simulated a dichotomous treatment as a function of the 100 baseline covariates. We then simulated a dichotomous outcome as a function of the 100 baseline covariates and treatment.

We considered four scenarios where we varied the sample size and the strength of the effects of covariates on both the treatment and outcome. We allowed coefficients to be both positive and negative to reflect practical settings where baseline covariates induce confounding in both directions. We implemented the PSs and DRSs using 1 to 1 caliper matching where calipers were defined as .01 standard deviations of the respective PS or DRS distribution. We compared the performance between PS and DRS matching by calculating the bias, defined as the expected value of the difference between the effect

estimate and the true effect, by taking the mean of this difference over all simulation runs. The mean squared error (MSE) was calculated by taking the mean of the squared bias over all simulation runs. We evaluated the precision of the effect estimates using the empirical standard deviation of the distribution of the treatment effect estimates across all simulation runs.

2.5.2 Simulations for Aim 2

We created a variety of populations where we varied the causal structures, covariate distributions and covariate associations. We evaluated scenarios which include rare outcomes, rare exposures and small sample sizes. In comparison to the complexities of real world data, these simulated populations were simplified in order to obtain a general understanding of the statistical properties and performance of various measures for evaluating the validity of DRS models. Simulations allow us to identify specific settings and parameters which systematically affect the performance of each of the described methods for confounding control.

2.6 Aim 1: Evaluating Potential Benefits of the Disease Risk Score

We will evaluate potential benefits of matching on the DRS compared to matching on the PS. Potential benefits of the DRS remain largely unclear with few studies demonstrating the application of DRSs in large database research.

2.6.1 Overlap in the distribution of disease risk across treatment.

Strong channeling can create separation in the PS distribution across treatment groups during early periods of drug approval. This can reduce the number of individuals who can be compared or matched on the PS and limit the ability to evaluate the treatment

effect within a large portion of treated patients, particularly when controlling for large numbers of covariates.

A theoretical advantage of the DRS that has not been discussed is that the degree of overlap in the DRS distributions across treatment groups will always be as at least as large as the overlap in the PS distributions across treatment groups. Greater overlap in the DRS distributions can potentially allow for a greater number of individuals to be compared when matching on the DRS compared to matching on the PS.

The reason for greater overlap in the distribution of disease risk across treatment is because of differences in the balancing properties of the PS compared with the DRS. Matching treatment groups on the PS is a more restrictive condition then matching treatment groups on the DRS. Matching on the DRS does not require covariates to be balanced across treatment groups and can include individuals who systematically have differing covariate distributions across treatment groups, but similar overall risk for the outcome.[5]

More formally, because matching on the PS renders baseline covariates independent of treatment assignment within the matched population, any function of baseline covariates will also be independent of treatment assignment including the DRS. Formally,

$$E[T|X, DRS] = E[T|X]$$

since the DRS is simply a function of X. Rosenbaum and Rubin show that

$$E[T|PS] = E[T|X],$$

implying that $E[T|X, DRS] = E[T|PS]$. In other words, once we condition on the PS, the covariate vector, X, or any function of X, including the DRS, does not provide any additional information about treatment assignment.

2.6.2. Controlling for instrumental variables

Including instrumental variables within the PS can also increase the separation in the PS distribution across treatment groups further limiting the number of patients that can be matched or compared. In addition, both theory and simulations have shown that controlling for variables that do not affect the outcome except through treatment (instrumental variables) can reduce the precision of effect estimates and amplify bias caused by unmeasured confounders.[62, 63]

Unmeasured confounding is a fundamental obstacle in pharmacoepidemiology and observational research in general. Primary sources of unmeasured confounding in comparative effectiveness research arise from 1) confounding by indication for treatment and 2) confounding caused by frailty.[57, 64, 65] For example, physicians' treatment decisions may be based on an evaluation of the patient's health status and prognosis, the patient's theoretical response to treatment, the physician's past experience with the treatment, or an assessment of the patient's ability and willingness to undergo the treatment (e.g., take a medication as prescribed).[66] Further, patients who initiate a preventive medication may be more likely than other patients to engage in other healthy, prevention-oriented behaviors leading to bias known as the healthy user effect.[67, 68] Conversely, patients in whom the expected benefit is unlikely to materialize (e.g., because of overwhelming competing risks) are less likely to be started on preventive therapies and more likely to stop preventive

therapies. These types of confounding usually result in the distributions of unmeasured covariates being systematically different between treatment groups.

For studies involving large numbers of covariates, identifying instrumental variables can be challenging. Pharmacoepidemiologic and medical studies utilizing automated databases often involve large numbers of potential covariates that have not been selected with a specific research question in mind and where a multitude of factors other than the prognosis strongly influence treatment decisions.

While the potential for including instrumental variables is highest for a model predicting treatment (the PS), it is important to realize that it is not generally avoided by modeling the risk for the outcome. Once we condition on treatment (either by modeling treatment or by restricting to the untreated), instrumental variables will become associated with the outcome via the unmeasured confounders. The DRS, if estimated within the study population, will also tend to be affected by bias amplification in the presence of unmeasured confounders.

By modeling covariate associations with the outcome within historical data prior to treatment introduction, the DRS implicitly avoids controlling for instrumental variables.[32] Out-of-sample estimation of DRSs is therefore likely to minimize bias caused by unmeasured confounding compared with PS methods or outcome regression models, including conventionally estimated DRSs. These potential advantages of greater overlap in the DRS distribution across treatment groups and the avoidance of controlling for instrumental variables can potentially improve the precision of effect estimates by allowing

for a larger proportion of treated individuals to be compared during periods where the number of individuals receiving the new treatment therapy can be small.

2.7 Aim 2: Model Validation

2.7.1. Validation of Propensity Score Models

We will evaluate the validity of the specified PS models by calculating the average standardized absolute mean difference (ASAMD) of the measured covariates across treatment groups, where the ASAMD for a single covariate is defined as

$$\frac{\frac{1}{n_1}\sum_{i=1}^{n_1} X_{i1} - \frac{1}{n_0}\sum_{i=1}^{n_0} X_{i0}}{\sqrt{(s_1^2 + s_0^2)/2}}$$

where

$X_{i1} =$ the value of X for the i^{th} individual in the treatment group,

$n_1 =$ number of individuals in the treatment group,

$s_1^2 =$ sample variance of X in the treatment group.

The ASAMD is a straightforward measure to summarize covariate balance and has been shown to perform well compared to other measures of covariate balance when evaluating the validity of PS models in their ability to control confounding.[69, 70]

2.7.2. Validation of Disease Risk Score Models

Hansen[5] showed that unlike propensity balance where covariates are balanced with respect to treatment within the entire study population, conditioning on the DRS results in 'prognostic balance' where covariates are balanced with respect to the potential outcome under the comparator treatment. Because this potential outcome is not observed for each individual in the study cohort, it is not possible to evaluate the validity of DRS models

directly in terms of prognostic balance across the entire population. It remains unclear what measures are optimal for evaluating risk models in their ability to control for confounding.

Within the simulated populations, we evaluated the correlation between various measures for assessing the validity of fitted DRS models and bias in the estimated treatment effect. We evaluated the predictive performance of the estimated DRSs by assessing the calibration and discrimination of the predicted values. The calibration was assessed using the Hosmer-Lemeshow goodness of fit test[71] defined as

$$H = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - E_g/n_g)} \sim \chi^2_{G-2}$$

where

$n_g =$ number of observations in the gth group,

$O_g = \sum_i y_{ig} =$ observed number of cases in the gth group,

$E_g = \sum_i p_{ig} =$ expected number of cases in the gth group.

We also evaluated the calibration of predicted values by calculating the prediction error for each DRS model. The mean prediction error was calculated as

$$PE = \frac{1}{n}\sum_i^n (y_i - p_i)^2,$$

where $y_i =$ the observed response for individual i, and p_i is the predicted response from the risk model.

We assessed the discrimination of the predicted values by calculating the c-statistic defined as the area under the receiver operating characteristic curve.[72] This curve is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) and is

calculated by comparing all disjoint pairwise combinations of individuals and calculating the proportion of those combinations where the predicted value for the individual with the event is higher than the predicted value for the individual with no event.

Finally, we evaluated the performance of a resampling method proposed by Hansen[73] that assesses the validity of DRS models in their ability to control for confounding rather than their predictive performance. This strategy draws weighted samples from the control population to create "pseudo treatment" and "pseudo control" groups so that the distribution of covariates across the pseudo treatment groups are representative of those across the treatment groups in the original study cohort. With no treatment effect separating the pseudo treated and pseudo control groups, the fitted DRS models can be evaluated directly in their ability to control for confounding within the pseudo population.

The pseudo treatment and pseudo control groups are created as follows:

1. Estimate the propensity score for each individual in the study population.
2. Create a pseudo treatment group by taking a weighted sample of $i = 1, \dots, k$ individuals from the control population with weights proportional to the odds of receiving treatment. Formally,

$$w_i \propto \frac{e(X)}{(1 - e(X))}$$

where $w_i =$ weight for individual i and $e_i(x) =$propensity score for individual i.

3. Sampling in step 2 can be done with or without replacement. When sampling with replacement, a pseudo control group is created by sampling $i = 1, \dots, m$

individuals from the control population with weights equal to one. If sampling without replacement, the pseudo control group is created by simply taking individuals from the control population that were not selected for the pseudo treatment group in step 2 (the complement of the pseudo treatment group).

Sampling the pseudo treatment group with replacement is similar in concept to standardized mortality ratio weighting where a subset of the control population is weighted so that the propensity score distributions between the weighted control group (which in this case is the pseudo treatment group) and treatment group in the original study cohort are exchangeable. Because $X \perp T|\psi(X)$, where X is a set of measured baseline covariates and $\psi(X)$ is the propensity score, this exchangeability on the PS implies exchangeability on the measured covariates X. With the covariate distributions across the pseudo treatment groups being representative of those within the actual treatment group, we can perform a "dry run" by fitting the DRS within the pseudo control group, or a historical set of controls, and then evaluating its ability of the fitted model to control for confounding within the full pseudo population. Since the degree of confounding within the pseudo population should be representative of confounding in the original study cohort, DRS models that successfully control for confounding within the pseudo population should also control for the same measured confounders within the original study cohort.

When sampling with replacement, the same individuals can potentially be sampled many times over. Sampling without replacement avoids this problem, but requires a more complicated sampling scheme to take into account that the sampling probabilities for each

individual change with each draw from the finite population. Chen & Dempster[74] describe a method to maximize information when taking a weighted sample without replacement from a finite population. In this paper, we restricted analyses to sampling with replacement.

CHAPTER 3

USING DISEASE RISK SCORES TO CONTROL LARGE NUMBERS OF COVARIATES IN COMPARATIVE EFFECTIVENESS RESEARCH OF NEW TREATMENTS

3.1 Introduction

Evaluating the comparative effectiveness of newly introduced treatments presents unique challenges in pharmacoepidemiologic research. The propensity score, defined as the conditional probability of treatment given a set of observed covariates, has become a standard tool for controlling large numbers of confounding variables.[1, 2] However, accurately modeling the PS for a new treatment can be difficult if the treated population is small or factors affecting treatment assignment change rapidly.[3, 4]

In a recent paper, Glynn et al.[3] proposed using an alternative covariate summary score, the disease risk score (DRS), to control for confounding in settings involving new treatments. Unlike the PS, which models covariate associations with treatment, the DRS models covariate associations with the outcome within the control or comparator treatment group. Glynn et al. argued that factors affecting disease risk are more likely to be stable over time than are factors affecting treatment, potentially simplifying the estimation of the DRS compared to a time-varying PS. Glynn et al. also advocated using data collected prior to the introduction of the new treatment to avoid overfitting and provide ample data when fitting rich prediction models.

Little evidence exists to confirm the theoretical advantage of a historically estimated DRS over a traditional PS when evaluating new treatments. A number of studies have shown that simply fitting time-specific PS models can perform well when the indication for treatment changes rapidly over time.[4, 75] Further, the limitations of the PS when the number of exposed individuals is small are not well understood. Previous studies have also shown that overfitting the PS model does not necessarily compromise confounding control.[76] There remain few examples demonstrating the application of a historically estimated DRS when evaluating new treatments. Potential advantages and challenges of using DRSs in these settings remain unclear.

In this paper, we use both simulations and an empirical example to compare the performance of the DRS with that of the PS when controlling large numbers of covariates in settings involving newly introduced treatments. We discuss both challenges and potential advantages of using the DRS for confounding control as well as required assumptions for using historical data to model the DRS. We then evaluate the performance of DRS matching with PS matching in an empirical example where we compare the new oral anticoagulant dabigatran with warfarin in preventing ischemic stroke and all-cause mortality in patients diagnosed with atrial fibrillation (AF) in the Medicare population.

3.2 Background

In comparative effectiveness research, investigators are often interested in comparing two alternative treatment therapies. Following Rubin's[77] description of the counterfactual framework, let Y_t represent the potential response had the individual received the treatment of interest and Y_c the potential response had the individual received

the comparator or control treatment. In practice, only one of these potential outcomes is observed for each individual.[18, 19]

Hansen[5] formally defines the DRS as any scalar or multidimensional function of a set of baseline covariates, X, that, when conditioned on (e.g., matching or subclassification), results in X being independent of Y_c. In the absence of unmeasured confounding, this independence is sufficient to identify average treatment effects in the treated population.[5, 21] If the outcome follows a generalized linear model, one possible DRS is the linear predictor of Y_c, or the conditional mean of Y_c given X (i.e., $E[Y_c|X]$). Because Y_c is observed only for individuals receiving the comparator treatment, in practice the DRS must be estimated indirectly for the treated population.

Challenges when modeling the DRS. The DRS has typically been estimated in two ways. The first is to fit a regression model within the cohort of individuals receiving the comparator treatment and then extrapolate this model to predict disease risk for the full cohort. The second is to fit a regression model for the full cohort (i.e., both treatment and comparator groups) as a function of baseline covariates and treatment, and then estimate the disease risk for each individual after setting treatment status to zero.[3, 5, 7, 27, 31] Fitting the DRS to the full cohort benefits from increased sample size, but requires accurately modeling the relation between treatment and outcome. Small misspecifications in the full-cohort DRS model can introduce bias by resulting in estimated scores that are non-ancillary, or carry information about the treatment effect.[5, 30] Consequently, Hansen[5] recommends using only the control population when fitting the DRS model. Leacy[30] explained that using only the control population when modeling the DRS tends to result in estimated scores that

are more robust to model misspecification. Fitting the DRS only among individuals receiving the comparator treatment, however, can lead to overfitting,[5] which results in overestimating disease risk for high-risk comparator patients and underestimating disease risk for low-risk comparator patients. Such overfitting can lead to apparent treatment effect heterogeneity over the distribution of disease risk and potentially bias overall effect estimates.[3, 5]

Both Hansen[5] and Glynn et al.[3] have proposed using controls from a period prior to the current study to fit the DRS model. Glynn et al. suggested that estimating the DRS with historical data can be particularly advantageous in pharmacoepidemiologic studies using large administrative healthcare databases to evaluate newly introduced treatments or evolving drug therapies. This approach can avoid overfitting the risk model to the comparator group within the study cohort, but assumes that the effects of risk factors on the outcome, surveillance of individuals, and coding practices do not change over time. Violation of these assumptions could result in fitted DRS models that are not generalizable to the study population.

Potential benefits of matching on the DRS. A theoretical advantage of the DRS that has not been widely discussed is that the degree of overlap in the distribution of disease risk across treatment groups will always be at least as large as the overlap between the PS distributions. This is due to the fact that matching on the PS is more restrictive than matching on the DRS. Matching on the PS will only include individuals who, in expectation, have similar covariate distributions across treatment. Matching on the DRS, however, will not only include individuals who, in expectation, have similar covariate distributions across

treatment, but can also include individuals who systematically have differing covariate distributions across treatment, but similar overall risk for the outcome.[5] More formally, once covariates are independent of treatment, then any function of baseline covariates, including the DRS, will also be independent of treatment. Therefore, PS-matched treatment groups will be balanced on the DRS in expectation. However, because the DRS does not balance covariates with respect to treatment, but only with respect to Y_c, DRS-matched groups may not be balanced on the PS. The potential for greater overlap in DRS distributions across treatment groups may allow a larger percentage of the treated population to be compared when matching on the DRS versus the PS.

3.3 Simulation Study: an Illustrative Example

We simulated a causal scenario that was motivated by an empirical example (described below) comparing dabigatran with warfarin in preventing ischemic stroke and all-cause mortality among new users. We simulated 100 baseline covariates. As in most pharmacoepidemiologic settings, the majority of these baseline covariates were dichotomous (simulated as binomial random variables). We simulated a dichotomous treatment and a dichotomous outcome according to equations 3.1 and 3.2.

$$logit(E[T|X_i]) = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_{100} X_{100} \quad [3.1]$$

$$logit(E[Y|X_i, T]) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{100} X_{100} + \beta_{int} X_1 T + \beta_T T \quad [3.2]$$

We considered four scenarios where we varied the sample size and the strengths of covariate-treatment and covariate-outcome associations. In scenario 1, coefficients in Equations 1 and 2 were selected randomly from uniform distributions so that the effects of covariates on both the treatment and outcome were mild. In Scenarios 2 and 3, coefficients

were selected to allow for moderate and strong effects, respectively, on the treatment and outcome. In Scenario 4 we included treatment effect heterogeneity to demonstrate how effect estimates can differ when different numbers of individuals are matched on the PS versus the DRS. We allowed coefficients to be both positive and negative to reflect practical settings where baseline covariates induce confounding in both directions. We simulated each scenario with sample sizes of 10,000 and 1,000.

We measured the performance of DRS and PS matching in three ways. We calculated the bias, defined as the expected value of the difference between the effect estimate and the true effect, by taking the mean of this difference over all simulation runs. The mean squared error (MSE) was calculated by taking the mean of the squared bias over all simulation runs. To evaluate precision, we estimated the standard error using the empirical standard deviation of the distribution of the treatment effect estimates across all simulation runs.

3.4 Simulation Results

For simulation scenarios not involving treatment effect heterogeneity, Figures 3.1 and 3.2 show the PS and DRS distributions by treatment group for one simulation run with a sample size of 10,000 (Figure 3.1) or 1,000 (Figure 3.2). As expected, the degree of overlap (i.e., area of overlapping region) between the DRS distributions was always larger than the degree of overlap between the PS distributions. Varying the sample size and the strengths of covariate-treatment and covariate-outcome associations affected the overlap in PS distributions more strongly than it affected the overlap in DRS distributions (Figures 3.1 and 3.2).

Table 3.1 shows that, for every scenario, a larger percentage of the treated population could be matched on the DRS versus the PS because of the greater overlap in DRS distributions (percent matched was approximately 100 for DRS matching and ranged from 98.8 to 47.6 for PS matching). The DRS-matched estimate generally had greater precision and lower MSE compared to the PS-matched estimate, with MSE ranging from 0.02 to 0.72 for DRS matching and 0.04 to 0.72 for PS matching (Table 3.1). Both DRS and PS matching resulted in approximately unbiased estimates for scenarios where there was no treatment effect heterogeneity. In the presence of treatment effect heterogeneity, matching on the DRS resulted in a more accurate evaluation of the treatment effect within the entire treated population (Table 3.1).

3.5 Empirical Study: Dabigatran vs Warfarin in Patients with Atrial Fibrillation

We compared the performance of dabigatran versus warfarin in an elderly population using linked Medicare Parts A (hospital), B (outpatient), and D (pharmacy) data. We identified eligible individuals from a 20% random sample of Medicare beneficiaries with fee-for-service enrollment in all three plans for at least one month from October 19, 2010 (when dabigatran was introduced) through December 31, 2012. New users were defined as individuals who initiated dabigatran or warfarin after a 1-year washout period with no prescription for any oral anti-coagulant.[11] We required continuous enrollment in Medicare for at least 12 months prior to drug initiation. All demographic and clinical covariates (described below) were defined during the 12 months prior to drug initiation. Individuals were censored only if they lost Medicare enrollment during follow-up (intent-to-treat analysis).

We restricted our study cohort to individuals who were 65 years of age or older and had an inpatient or outpatient diagnosis code for atrial fibrillation or atrial flutter (ICD-9 427.31, 427.32) prior to initiation of dabigatran or warfarin. We excluded individuals with a known heart valve replacement because this is a contraindication for dabigatran use. We also excluded individuals at a skilled nursing facility at drug initiation.

Estimation of the DRS and PS. We modeled the one-year risk of combined ischemic stroke and all-cause mortality within a population of new warfarin users with an index date prior to the introduction of dabigatran (between January 1, 2008 and October 18, 2010). This model was then used to predict the disease risk for all individuals within the study cohort.

We also estimated the PS within the study cohort for comparison. The PS and DRS models included main effects for the 37 covariates listed in Table 3.2, which were selected a priori using expert knowledge. We added 200 empirically selected covariates based on Medicare medication claims, inpatient and outpatient diagnostic codes, and procedural codes. We identified the 200 most prevalent codes within each data dimension (codes with a prevalence greater than 0.5 were subtracted from 1). Of the 600 covariates identified in this way, we selected the 200 with the strongest univariate associations (odds ratios) with the outcome. The estimated DRS and PS were implemented using 1-1 caliper matching with a caliper of 0.01 standard deviations of the respective DRS or PS distribution. We considered the dabigatran group to be the treated group. We estimated the hazard ratio within the matched populations using Cox proportional hazards models.

We conducted analyses using 20 and 1 percent samples of the Medicare data to

observe the sensitivity of the results as the sample size is reduced. With smaller sample sizes, PS overfitting and the resulting separation in PS distributions across treatment groups will likely be more pronounced. Previous studies have shown that confounding can be stronger shortly after a treatment's introduction.[78-80] To observe the sensitivity of the results to the duration of follow-up, we repeated the analysis using data only for the first year of dabigatran use (index date between October 19, 2010 through October 18, 2011).

3.6 Empirical Results

We present results for the empirical study in Figures 3.3 and 3.4 as well as Tables 3.2 and 3.3. Table 3.2 shows the distribution of the 37 a priori selected covariates by treatment group. New users of dabigatran were generally healthier, with fewer comorbidities and greater use of the healthcare system than new users of warfarin (Table 3.2). Similar patterns of initiation have been found in other studies.[81]

Figures 3.3 and 3.4 show the PS and DRS distributions by treatment group for the 20% (Figure 3.3) and 1% (Figure 3.4) samples of the Medicare data, with follow-up through 2012. In both analyses, controlling for the larger set of empirically selected covariates resulted in greater separation in PS distributions while having little impact on the separation in DRS distributions. For the 20% sample (Table 3.3), approximately 100% of the treated population was matched on the PS and the DRS, regardless of the number of covariates included in the models. In this case, both PS and DRS matching resulted in similar hazard ratios and standard errors, both when controlling for the covariates selected a priori (HRs 0.73 and 0.72 respectively; SEs both 0.03) and after adding the empirically selected covariates (HRs 0.88 and 0.87 respectively; SEs both 0.04).

When using the 1% sample of the Medicare data and controlling for the covariates selected a priori (Table 3.4), PS and DRS matching yielded similar results, with approximately 100% of the treated population being matched for both methods (HR and SE of 0.75 and 0.16 for PS matching and 0.75 and 0.15 for DRS matching). However, when controlling for the expanded covariate set in this sample, only 82% of the treated patients were matched on the PS, compared to approximately 100% on the DRS (Table 3.4). The reduction in the percentage matched resulted in reduced precision for the PS-matched estimate (SE 0.21 versus 0.18) (Table 3.3). In the analyses evaluating treatment effects in the first year of dabigatran use (not shown), the pattern of results was similar to that shown in Table 3.3, except that unadjusted and adjusted estimates were further from the null and standard errors were larger.

Each of the PS models resulted in good model fit in terms of calibration and discrimination for all scenarios (Hosmer-Lemeshow p-value ranging from 0.52 to 0.65; c-statistic ranging from 0.68 to 0.79). The PS models also performed well in terms of balancing covariates across treatment groups with an average standardized absolute mean difference (ASAMD) of 0.01 or less for all scenarios. In terms of predictive performance, the DRS models had good discrimination (c-statistic ranging from 0.73 to 0.78), but performed poorly in terms of calibration (Hosmer-Lemeshow p-value <0.01 for three out of four scenarios).

3.7 Discussion

In this study, we used both simulations and an empirical example to explore potential benefits of using a historically estimated DRS when controlling large numbers of

covariates in settings with newly introduced treatments. With few exposed individuals and smaller sample sizes, fitting a high-dimensional PS model can increase separation between the PS distributions of the treatment groups, reducing the number of treated individuals who can be matched on the PS. In theory, the overlap in DRS distributions should always be at least as great as the overlap in PS distributions across treatment groups. Therefore, matching on the DRS may allow researchers to evaluate the treatment effect within a larger proportion of the treated population, compared to matching on the PS.

In the simulations, we demonstrated that when there was strong separation in the PS distributions across treatment groups, matching on the DRS can improve the precision of the effect estimate and, in the presence of treatment effect heterogeneity, provide more accurate estimates of the treatment effect in the full treated population. For the empirical example, we found that when there was moderate separation in the PS distributions across treatment groups, DRS and PS matching gave similar estimates of the effect of the new oral anticoagulant dabigatran versus warfarin in reducing combined ischemic stroke and all-cause mortality within the Medicare population. However, when controlling for large numbers of covariates with reduced sample size, the separation in the PS distribution across treatment groups increased and matching on a historically estimated DRS improved the precision of the effect estimate by allowing a larger proportion of the treated population to be matched. For both PS and DRS matching, when we added a large set of empirically selected covariates, effect estimates became more consistent with the results of clinical trials and other studies comparing these treatments within the Medicare population.[49, 82] When we restricted the analysis to the first year of dabigatran use, estimates moved further

from the null (becoming less consistent with trial results), likely reflecting the strong channeling that occurs shortly after a treatment's introduction.[78-80]

While matching on the DRS can allow for a larger portion of individuals to be compared across treatment when there is separation in the PS distributions, it is important to consider why the PS distributions are separated. If the separation is due to strong differences in confounding variables rather than overfitting the PS model, researchers should proceed cautiously. Strong differences in measured confounders can indicate strong differences in unmeasured confounders, which could be addressed best in the study design phase rather than the analysis phase. We stress the importance of reducing differences in the distribution of baseline covariates across treatment groups through proper study design (e.g., new-user design and other restriction criteria).[78, 83]

While we have focused on potential benefits of matching on the DRS, the DRS also has some theoretical disadvantages compared to the PS. Because the DRS is defined in terms of a potential outcome, estimating the DRS in practice can be challenging and requires additional assumptions. Further, unlike the PS, the DRS cannot be evaluated using measures of covariate balance within the full study population. In this study, the estimated PS models resulted in good model fit and PS matching balanced covariates across treatment groups. When modeling the DRS using historical data, we found it difficult to obtain good model fit in terms of the Hosmer-Lemeshow test, particularly when controlling for larger numbers of covariates. Other studies have reported similar findings when estimating high-dimensional DRSs and have proposed implementing shrinkage methods to reduce the dimensionality of covariates to improve model fit.[84] For this study, however, poor fit in

terms of the Hosmer-Lemeshow test did not appear to have a strong impact on the performance of the DRS compared with the PS. More research is needed to determine how best to estimate and evaluate the validity of DRS models.

We conclude that under certain assumptions, using historical data to model the DRS is a valid method to control for confounding when evaluating newly marketed drugs. Further, when there is strong separation in the distribution of the PS across treatment groups, matching on a historically estimated DRS versus a PS can allow researchers to evaluate the treatment effect within a larger proportion of the treated population. We further conclude, however, that accurately modeling the DRS can be more challenging as compared to modeling the PS, even in settings involving newly introduced treatments. When using summary scores for confounding control, we recommend conducting and reporting results from PS analyses in addition to analyses using a historically estimated DRS.

3.8 Tables and Figures

Table 3.1 Simulation results

Scenario[a]	Sample Size	Method	Bias	St. Error	MSE x10	% matched
A						
	10,000					
		Unadjusted	0.06	0.08	0.11	-----
		PS match	0.01	0.09	0.08	98.8
		DRS match	0.01	0.09	0.08	99.9
	1,000					
		Unadjusted	0.06	0.23	0.56	-----
		PS match	0.00	0.27	0.72	89.2
		DRS match	0.01	0.27	0.72	99.8
B						
	10,000					
		Unadjusted	0.12	0.06	0.18	-----
		PS match	0.00	0.08	0.06	89.9
		DRS match	0.01	0.07	0.06	99.9
	1,000					
		Unadjusted	0.13	0.22	0.65	-----
		PS match	0.00	0.26	0.67	77.4
		DRS match	0.00	0.22	0.48	99.9
C						
	10,000					
		Unadjusted	0.23	0.05	0.55	-----
		PS match	0.00	0.06	0.04	58.9
		DRS match	0.00	0.04	0.02	100
	1,000					
		Unadjusted	0.23	0.16	0.79	-----
		PS match	0.01	0.25	0.63	47.7
		DRS match	0.01	0.15	0.23	99.8
D						
	10,000					
		Unadjusted	0.22	0.05	0.51	-----
		PS match	0.04	0.06	0.05	58.9
		DRS match	0.01	0.04	0.02	100
	1,000					
		Unadjusted	0.22	0.18	0.81	-----
		PS match	0.05	0.24	0.60	47.6
		DRS match	0.01	0.16	0.26	99.9

[a] Scenario A: mild covariate effects on treatment and outcome; Scenario B: moderate covariate effects on treatment and outome; Scenario C: strong covariate effects on treatment and outcome; Scenario D: treatment effect heterogeneity with strong covariate effects on treatment and outcome

Table 3.2 Baseline covariates measured during 1-year washout period

	Warfarin (N=56,260)	Dabigatran 150mg (N=11,407)
Demographics:		
Age	78.91	76.76
Race (1 white, 0 other) (%)	89.2	91.72
Sex (% female)	42.17	48.95
% Diagnoses		
Cardiovascular:		
Chest pain	38.41	35.05
Heart disease	74.56	66.62
Heart failure	30.74	19.23
Hypertension	65.08	63.30
Hyperlipidemia	35.21	41.09
Myocardial Infarction	3.49	1.89
Cerebrovascular disease	21.29	17.38
Stroke		
Ischemic	6.09	4.31
Hemorrhagic	0.34	0.16
TIA	6.9	6.34
VTE	10.36	1.67
Diabetes	35.09	30.02
Kidney disease	12.58	4.74
Renal failure	16.09	5.75
Bleeding	1.88	0.68
Anemia	15.63	9.95
Baseline Meds: (%)		
Anti-depressants	28.27	22.89
Antihypertensives:		
ACE/ARB	52.22	50.23
Loop diuretics	40.91	28.70
Nonloop diuretics	52.55	41.97
Hypolipidemic drugs:		
Statins	49.40	52.45
Fibrate	5.02	4.98
Rate Control Therapy:		
Beta blockers	70.83	71.99
CCB	43.97	41.80
Glycoside	18.49	17.10
Rhythm Control Therapy	19.10	23.21
Healthcare Use (average #):		
# ECG claims	3.74	3.80
# PSA claims	0.36	0.49
# of fecal occult blood tests	0.12	0.13
# colonoscopies	0.14	0.14
# flu shot claims	0.76	0.79
# of lipid assessments	1.52	1.72
# of mammography claims	0.25	0.29
# of PapSmear claims	0.05	0.07

Table 3.3 Empirical results comparing new users of dabigatran with new users of warfarin in preventing combined ischemic stroke and all-cause mortality in the Medicare population between October 19, 2010 and December 31, 2012.

Sample Size[a]	# covs[b]	Method	Hazard Ratio[c]	St. Error[d]	95% CI	% matched	Model Fit[e]		
							c-stat	p-value	ASAMD[f]
20% Sample									
		Unadjusted	0.48	0.02	(0.46, 0.50)	------	------	------	0.14
	37								
		PS match	0.73	0.03	(0.69, 0.77)	100	0.68	0.63	<0.01
		DRS match	0.72	0.03	(0.68, 0.76)	100	0.73	<0.01	------
	237								
		PS match	0.88	0.04	(0.81, 0.95)	100	0.73	0.52	<0.01
		DRS match	0.87	0.04	(0.80, 0.94)	100	0.78	<0.01	------
1% Sample									
		Unadjusted	0.47	0.07	(0.41, 0.54)				0.17
	37								
		PS match	0.75	0.16	(0.55, 1.03)	98.3	0.71	0.65	0.01
		DRS match	0.75	0.15	(0.56, 1.01)	100	0.73	0.18	------
	237								
		PS match	0.89	0.21	(0.59, 1.34)	81.5	0.79	0.61	0.01
		DRS match	0.86	0.18	(0.60, 1.22)	99.3	0.77	<0.01	------

[a] 20% (N=67,667) and 1% (N=3,383) samples of the Medicare data.

[b] Number of covariates in PS and DRS model

[c] RELY trial relative risk for 150mg dabigatran vs warfarin: 0.76 (0.60, 0.98) for ischemic stroke; 0.88 (0.77, 1.00) for death from any cause. In the current study, >90% of the outcomes were death from any cause.

[d] Bootstrapped standard errors. Hazard ratio estimates are the mean of the bootstrapped sampling distribution

[e] c-statistic and p-value for each PS and DRS model.

[f] The average standardized absolute difference (ASAMD) of covariates across PS matched treatment groups. Because the DRS does not balance covariates across treatment, the ASAMD was only calculated for PS models. The unadjusted ASAMD was calculated for all 237 covariates.

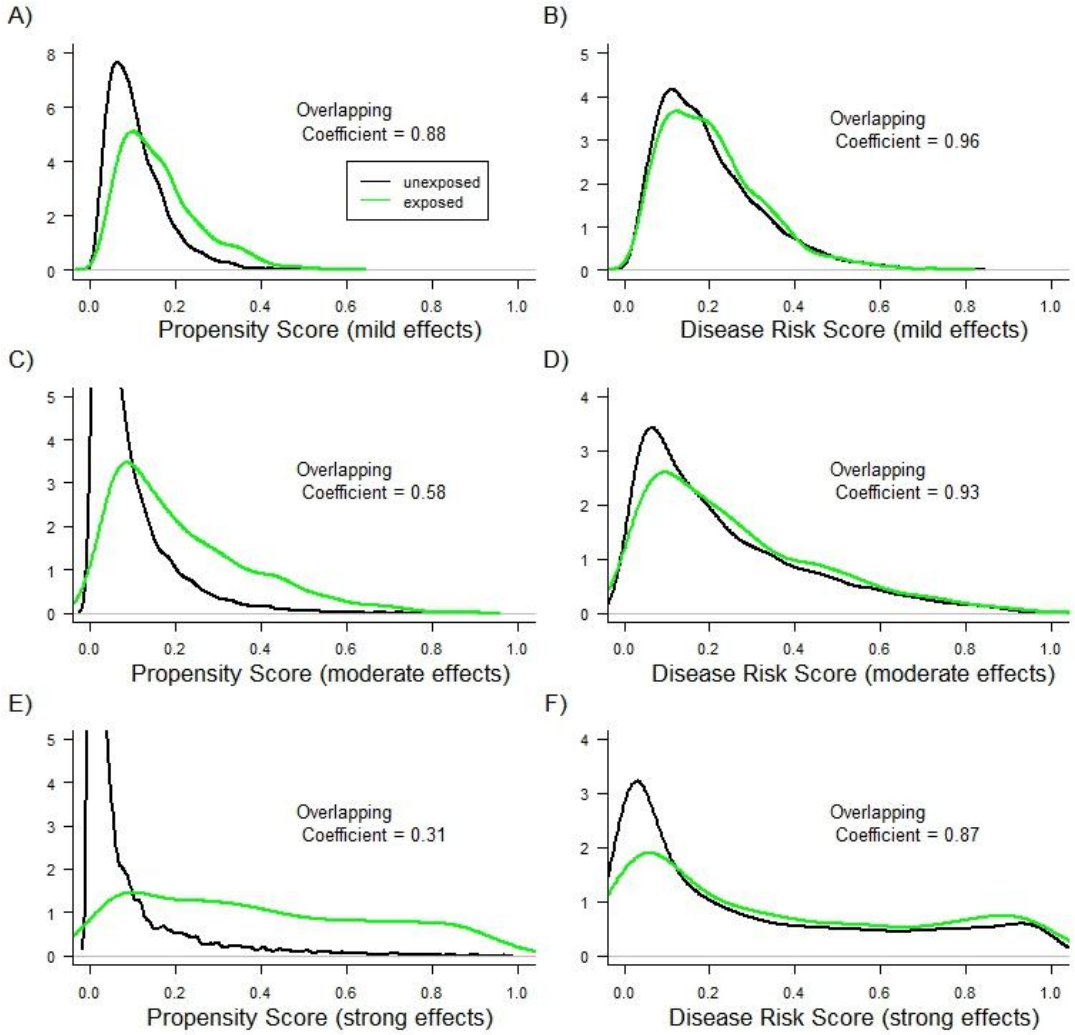

Figure 3.1 PS and DRS distributions across treatment groups with a sample size of 10,000 subjects and 100 covariates included in the PS and DRS models. In plots A and B the effects of covariates on both treatment and the outcome were mild, in plots C and D covariate effects were moderate, and in plots E and F the covariate effects were strong. The overlapping coefficient is an estimate of the percentage of overlapping area between the two density functions.

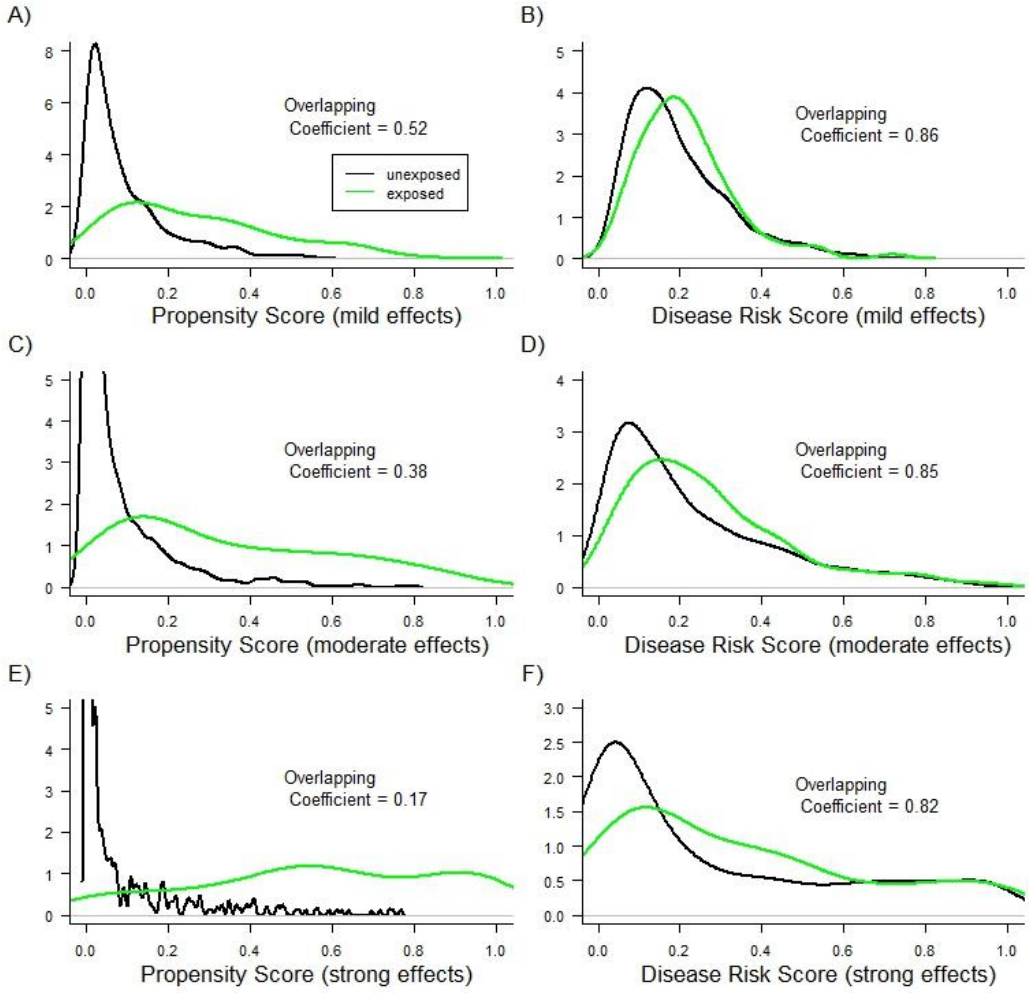

Figure 3.2. PS and DRS distributions across treatment groups for one run of the simulation study with a sample size of 1,000 subjects and 100 covariates included in the PS and DRS models. In plots A and B the effects of covariates on both treatment and the outcome were mild, in plots C and D covariate effects were moderate, and in plots E and F the covariate effects were strong. The overlapping coefficient is an estimate of the percentage of overlapping area between the two density functions.

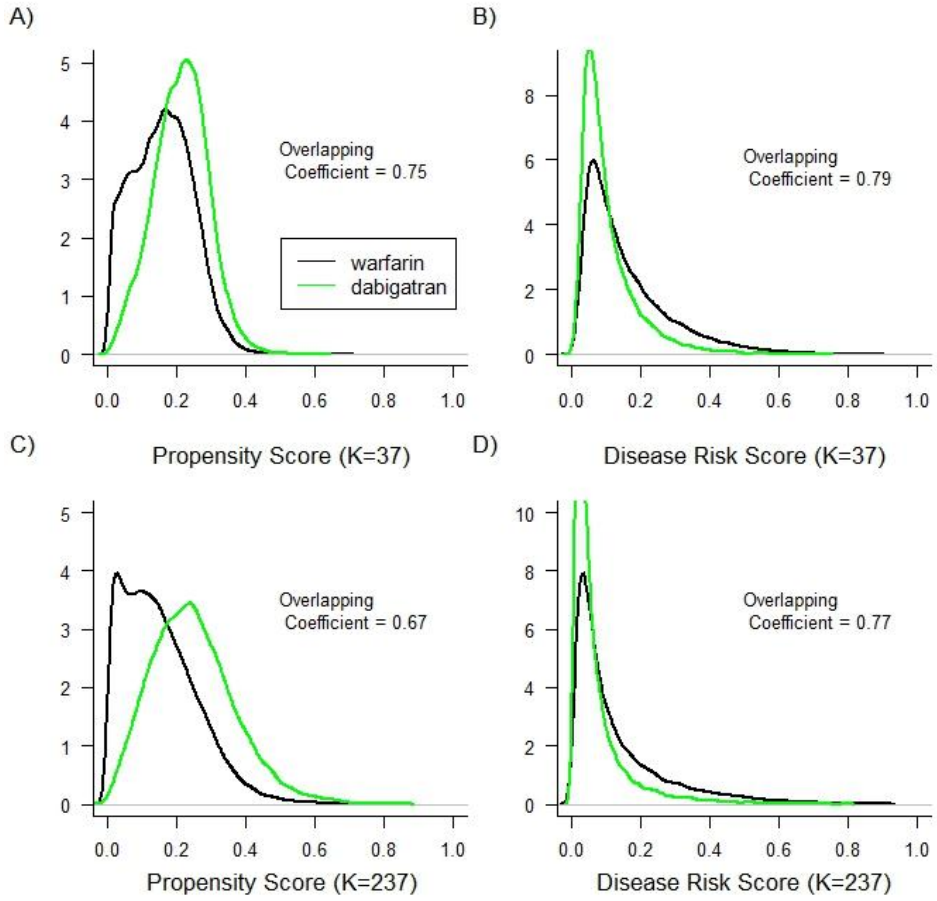

Figure 3.3. PS and DRS distributions across dabigatran and warfarin treatment groups for a 10 percent sample of the Medicare dataand individuals with an index date between October 2010 and December 2012. In plots A and B the PS and DRS models included 37 a priori selected covariates. In plots C and D the PS and DRS models included 37 a priori selected covariates and 200 empirically selected covariates. The overlapping coefficient is an estimate of the percentage of overlapping area between the two density functions.

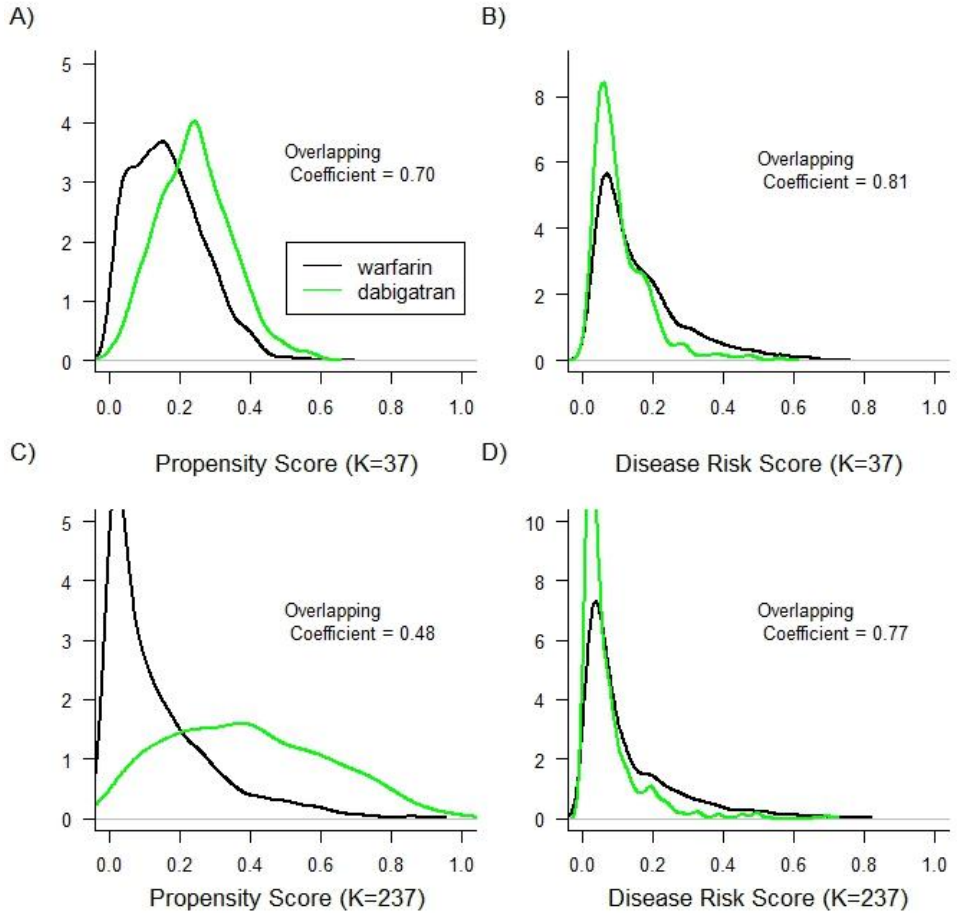

Figure 3.4. PS and DRS distributions across dabigatran and warfarin treatment groups for a 1 percent sample of the Medicare data and individuals with an index date between October 2010 and December 2012. In plots A and B the PS and DRS models included 37 a priori selected covariates. In plots C and D the PS and DRS models included 37 a priori selected covariates and 200 empirically selected covariates. The overlapping coefficient is an estimate of the percentage of overlapping area between the two density functions.

CHAPTER 4

METRICS TO EVALUATE DISEASE RISK SCORES IN NON-EXPERIMENTAL RESEARCH

4.1 Introduction

Controlling large numbers of confounding variables is a fundamental challenge in pharmacoepidemiologic research. Summary scores, which reduce baseline covariate information to a single dimension, have become increasingly popular for confounding adjustment. The propensity score, defined as the conditional probability of treatment given a set of measured covariates, has been the most widely used summary score for confounding control.[2] An alternative to the propensity score is the prognostic score, also known as the disease risk score (DRS). Unlike the PS which models covariate associations with treatment assignment, the DRS models the associations between covariates with the outcome under the control or comparator treatment.[5]

Both the PS and DRS control for measured confounders by acting as balancing scores. Rosenbaum & Rubin[1] show that upon conditioning on the PS, measured covariates are independent of treatment assignment. Hansen[5] shows that the DRS is also a balancing score, but instead of balancing covariates with respect to treatment assignment, the DRS acts as a "prognostic balancing" score in that conditioning on the DRS results in covariates being independent of the potential outcome under control (discussed further below).

In practice, the PS and DRS are unknown and must be estimated from the available data. A favorable aspect of the PS is the ability to check the validity of the model by

assessing the balance of covariates across treatment groups within the entire study population. A number of studies have shown a strong correspondence between measures of covariate balance and the ability of the PS model to reduce confounding bias.[85, 86] It has become common practice and recommended that PS estimation should be approached with the primary goal of minimizing covariate imbalance rather than focusing on the prediction of treatment assignment.[34, 87, 88]

The optimal strategy for evaluating DRS models is less clear. The goal of the DRS is to control confounding by balancing covariates with respect to the potential outcome under control. Because this potential outcome is only observed for individuals receiving the comparator treatment, "prognostic balance" cannot be evaluated within the entire study population. Consequently, the validity of fitted DRS models has primarily been assessed through measures of predictive performance, including the c-statistic and goodness of fit tests. It remains unclear how well measures of predictive performance correspond with the ability of the DRS to reduce confounding bias.

In this study we use simulations and an empirical example to compare metrics for evaluating the predictive performance of DRS models and their correspondence with bias in treatment effect estimates. We further discuss and evaluate a strategy proposed by Hansen[73] where DRS models are evaluated in their ability to control confounding by performing a "dry run" analysis within a pseudo population of individuals who are sampled from the control population in a way to represent the covariate distributions within the full study cohort. Finally, we demonstrate the discussed concepts through an empirical example comparing dabigatran vs warfarin in preventing ischemic stroke and all-cause mortality

within the Medicare population. While a number of studies have discussed and analyzed ways for evaluating fitted PS models,[85, 86] there remains little discussion in the literature regarding how DRS models should be evaluated when the goal of the DRS is to control for confounding bias.

4.2 Background and Notation

Following the Neyman-Rubin counterfactual framework,[18, 19] let Y_1 represent the potential response had the individual received treatment and Y_0 the potential response had the individual received the control or comparator treatment. In practice, only one of the potential outcomes is observed. Let Y represent the observed response and T a dichotomous treatment. Further, let X represent a set of measured baseline covariates with $\varphi(X)$ the PS and $\psi(X)$ the DRS, both as a function of the baseline covariates X.

Rosenbaum & Rubin[1] show that upon conditioning on the PS, measured covariates are independent of treatment assignment. Formally expressed as $X \perp T|\varphi(X)$ where $\perp$ denotes independence of random variables and | denotes conditional on, this independence results in covariates being balanced across treatment groups. If treatment assignment is strongly ignorable, i.e. $(Y_1, Y_0) \perp T|X$, this property of covariate balance that results from conditioning on the propensity score satisfies $(Y_1, Y_0) \perp T|\varphi(X)$ allowing for the identification of average treatment effects in the full population or within subgroups of the population (e.g., the treated population). If treatment assignment is weakly ignorable, i.e. $Y_0 \perp T|X$, then conditioning on the propensity score satisfies $Y_0 \perp T|\varphi(X)$ allowing for the identification of the average treatment effects on the treated population.

Formally, a prognostic score is defined as any scalar or multi-dimensional function of X that satisfies the condition $Y_0 \perp X|\psi(X)$.[5] In other words, conditioning on the DRS results in a form of covariate balance where the potential response under the comparator treatment is independent of a set of measured covariates, X. Hansen[5] shows that if treatment assignment is weakly ignorable, then conditioning on the DRS satisfies $Y_0 \perp T|\psi(X)$ allowing for unconfounded estimates of treatment effects in the treated population through stratification or matching.

If the outcome follows a generalized linear model, then one possible DRS is the linear predictor of Y_0, or the conditional mean of Y_0 given X (i.e., $E[Y_0|X]$). In practice, Y_0 is only observed for individuals receiving the comparator treatment and the function $E[Y_0|X]$ must be estimated indirectly for the treated population.

4.2.1 Challenges when modeling the DRS and evaluating its ability to control confounding

The DRS has primarily been estimated in one of two ways: 1) model $E[Y_0|X]$ within the cohort of individuals receiving the comparator treatment and then extrapolate this model to predict disease risk for all individuals within the full cohort; 2) model the function $E[Y|X,T]$ by fitting a regression model to the full cohort (i.e., both treated and control) as a function of baseline covariates and treatment, and then estimate $E[Y_0|X]$ for each individual through the function $E[Y|X,T=0]$ by setting treatment status to zero.[6, 7, 23, 30, 31]

Fitting the DRS to the full cohort benefits from increased sample size, but requires accurately modeling the relationship between the treatment and outcome. Small misspecifications in the full cohort DRS model can result in the estimated scores that carry information about the treatment effect. This non-ancillarity in the estimated DRSs can

introduce bias when used for confounding control.[5, 30] Consequently, Hansen[5] recommends using only the control population when fitting the DRS model to ensure partial ancillarity in the estimated scores. Leacy[30] explains that using only the control population when modeling the DRS tends to result in estimated scores that are more robust to model misspecification. Fitting the DRS only within individuals receiving the comparator treatment, however, increases the potential for overfitting the risk model to the control population in the study cohort. Such overfitting can overestimate disease risk for high-risk comparator patients and underestimating disease risk for low-risk comparator patients, leading to apparent treatment effect heterogeneity over the distribution of disease risk and potentially biased effect estimates.[3, 5, 32]

Both Hansen[5] and Glynn[3] have proposed that fitting the DRS within a historical sample of controls can reduce problems of overfitting. This strategy has received particular attention in pharmacoepidemiologic studies involving large automated databases and the evaluation of newly introduced treatments.[3, 32] Fitting the DRS within historical controls, however, requires additional assumptions that the effects of risk factors on the outcome, coding practices, and surveillance of individuals do not change over time. These difficulties when modeling the DRS highlight the importance of evaluating the validity of fitted DRS models.

If measures of prognostic balance were available within the full study population, these measures could be used to evaluate DRS models in a similar way that measures of covariate balance across treatment groups are used to evaluate PS models. However, prognostic balance can only be evaluated within individuals receiving the comparator

treatment. Measuring prognostic balance only within the comparator group can potentially reward models that are overfit to the control population. Further, in the presence of unmeasured confounding, the DRS does not result in prognostic balance within subgroups of treatment, but only marginally within the entire population.[5, 32] Fitted risk models that induce prognostic balance within the control population does not imply that those models will induce prognostic balance within the entire study cohort.

4.3 Dry Run Analysis

Hansen[73] proposes incorporating the propensity score when evaluating the ability of risk scores to control confounding. Hansen explains that, in theory, researchers can use the propensity score to draw weighted samples from the control population to create "pseudo treatment" and "pseudo control" groups whose covariate distributions on measured covariates are representative of the treated and control populations in the full study cohort.[73] With no treatment effect separating the pseudo treated and pseudo control groups, researchers can perform a "dry run" analysis by fitting the DRS model to the pseudo control group, or a historical set of controls, and then evaluate the validity of the fitted model based on its ability to control for confounding within the pseudo population. If subclassification or matching on the estimated DRSs result in unconfounded null effect estimates within the pseudo population, then the fitted model should be successful in controlling confounding on the same measured covariates within the original sample.

To create a "pseudo treatment" group, individuals are sampled from the control population within the study cohort with weights proportional to the odds of receiving treatment. For the "pseudo control" group, individuals are sampled with weights of one.

Sampling can be done with or without replacement. Sampling with replacement is similar in concept to standardized mortality ratio (SMR) weighting where a portion of the control population is weighted to represent the PS distribution within the treated population.[89] With smaller sample sizes or large weights, sampling with replacement can result in the same individuals being sampled many times over. Sampling without replacement avoids this problem, but requires a more complicated sampling scheme known as maximum entropy sampling to maximize information when sampling from a finite population. Chen et al.[74] provide a detailed explanation of maximum entropy sampling from finite samples.

In theory, this strategy provides a more direct approach for evaluating the ability of a DRS model to control for measured confounding. There remains little evidence, however, of its performance in practice and the optimal strategy for evaluating DRS models remains unclear.

4.4 Simulation Study

We simulated a dichotomous treatment (T) and outcome (Y), six binary covariates ($X_1, X_3, X_5, X_6, X_8, X_{10}$) and four standard-normal covariates (X_2, X_4, X_7, X_9). We defined the conditional probability of treatment and outcome according to Equations 4.1 and 4.2.

$$logit(E[T|X_i]) = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{10} X_{10} + \alpha_{11} X_2 X_3 + \alpha_{12} X_4 X_8 + \alpha_{13} X_5 X_6 + \alpha_{14} X_7^2 + \alpha_{15} X_{10}^2 \quad [4.1]$$

$$logit(E[Y|X_i, T]) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{10} X_{10} + \beta_{11} X_2 X_3 + \beta_{12} X_4 X_8 + \beta_{13} X_5 X_6 + \beta_{14} X_7^2 + \beta_{15} X_{10}^2 + \beta_T T \quad [4.2]$$

The coefficient values for α_i and β_i, i =1,…,10, were selected by drawing values from separate uniform(-.7,.7) distributions. This range of values (i.e., potentially ranging

from -0.7 to 0.7) was chosen to reflect the range for the majority of coefficient values observed in an empirical example comparing dabigatran versus warfarin described below. We repeated this process 50 times by drawing a separate set of values for α_i and β_i, i =1,…,10, to consider a total of 50 unique parameter combinations. With each parameter combination we simulated 100 datasets and fit 32 unique DRS models using logistic regression with various degrees of model misspecification. Each of the models included main effects for the covariates X_1 through X_{10}, but different sets of higher order terms. We considered all possible combinations of the higher order terms (32 total). We fit the DRS models within a simulated historical set of data that was similar to the study cohort, but with no treatment introduced.

We evaluated the predictive performance for each DRS model within the original study cohort by calculating three measures of predictive performance: the c-statistic, the Hosmer-Lemeshow goodness of fit test and the mean squared error (MSE) of the predicted values. We also evaluated each DRS model by performing a "dry run" analysis to calculate the pseudo bias after stratifying on the fitted DRSs within a pseudo population. We created pseudo treatment and control groups by sampling with replacement from the actual control population in the original study cohort with weights proportional to the odds of the PS or one respectively. Because the dry run analysis relies on using the PS to create the sampling weights, we estimated the PS using logistic regression and considered different degrees of misspecification: no misspecification (included all higher order terms), moderate misspecification (excluded one quadratic and one interaction term), and strong misspecification (excluded all higher order terms). Finally, we evaluated the correlation

between bias in the effect estimates with the measures of predictive performance and the calculated pseudo bias.

For each parameter combination, we repeated the analysis to consider scenarios involving smaller sample sizes, low prevalence of both treatment and outcome, risk factors (i.e., variables that affect the outcome but have no effect on treatment), different distributions for baseline covariates, and different directions in confounding. We also repeated all analyses using a probit model to simulate treatment and the outcome (i.e., probit rather than logit model in Equations 1 and 2). Simulating data with a probit model allowed us to evaluate scenarios where model misspecification in the PS and DRS models was a result of the functional form of covariates in the model as well as the functional form of the model itself (i.e., misspecified link function). In this case, there was mild misspecification in both the fitted PS and DRS models even when all higher order terms were included within the models. Table 4.1 provides a description of the simulated scenarios.

4.5 Simulation Results

We present results for the simulation study in Figures 4.1 and 4.2, and Table 4.2. Figures 4.1a through 4.1f show box plots for the Spearman correlation coefficients between each of the described measures for evaluating DRS models and the absolute bias in the treatment effect. Each box-plot shows the distribution of 50 correlation coefficients (one correlation coefficient for each of the 50 parameter combinations). Each sub figure represents a different simulation scenario.

When the PS was correctly specified, there was a very strong correlation between the calculated pseudo bias and the actual bias within the original study cohort (Figure 4.1a). As the misspecification in the PS model increased, the strength of this correlation became less pronounced (Figures 4.1c and 4.1e). The correlation between the measures of predictive performance and bias were less consistent and attenuated compared to the correlation between the full cohort study bias and the pseudo bias when the PS was correctly specified (Figures 4.1a-4.1f). Among the measures of predictive performance, the p-value for the Hosmer-Lemeshow goodness-of-fit test showed the weakest correlation with predicting bias in the effect estimate (Figure 4.1d). Similar patterns were found when treatment and outcome were simulated using a probit model (results not shown). In this case, pseudo bias1 represented the bias within the pseudo population when there was mild misspecification in the PS model.

Figure 4.1 shows that the c-statistic and MSE had good performance in predicting bias in the effect estimate when comparing various models fit to the same dataset. In practice, however, researchers will often want to make a decision on a single fitted model to decide if the model is appropriate. The calculated pseudo bias and p-value from goodness-of-fit tests may have an advantage in this case as the actual values from these measures can carry the same meaning across different data generating models (e.g., a p-value of <0.01 implies poor model fit regardless of the data generating model). To explore this issue, we calculated a single correlation for each measure after combining the calculated values across all parameter combinations and scenarios instead of calculating the correlation between each measure and bias separately for each data generating model

(Figure 4.2). Similar to Figure 4.1, when the PS was correctly specified, the calculated pseudo bias showed the strongest correlation with bias in the treatment effect and had a calculated intercept from the least squares regression line of approximately 0 (Figure 4.2a). Among the measures of predictive performance, the p-value from the Hosmer-Lemeshow test showed the strongest correlation after combining results across different data generating models (Figure 4.2d).

4.6 Empirical Example: Dabigatran vs Warfarin

We compared the performance of dabigatran vs warfarin in an elderly population using a 20 percent sample of linked Medicare parts A (hospital), B (outpatient), and D (pharmacy) data. We included Medicare beneficiaries with fee for service enrollment in all three plans for at least one month between October 19, 2010 through December 31, 2012. New users were defined as individuals who initiated dabigatran or warfarin after a 1 year washout period of having no prescription for either dabigatran, warfarin, or any other oral anti-coagulant. We included all individuals who were continuously enrolled in Medicare for at least 12 months prior to drug initiation. All demographic and clinical covariates were defined during the 12 months prior to drug initiation. Individuals were censored only if they lost enrollment in the Medicare system during follow-up (intent to treat analysis).

We restricted our study cohort to individuals who were 65 years of age or older and had an inpatient or outpatient diagnosis code for atrial fibrillation or atrial flutter (ICD-9 427.31, 427.32) prior to initiation of dabigatran or warfarin. We excluded individuals with a known heart valve transplant since this is a contraindication for dabigatran use. We also excluded individuals at a skilled nursing facility at drug initiation since diagnoses within

these facilities are not always captured within Medicare claims data.

To avoid overfitting the risk model to the control population within the study cohort, we modeled the one year risk of combined ischemic stroke and all-cause mortality within a historical population of new warfarin users with an index date prior to the introduction of dabigatran (between January 1, 2008 through October 18, 2010). This model was then used to predict the disease risk for all individuals within the study cohort. We fit PS and DRS models that included main effects for each of the covariates listed in Table 4.2 and an additional 200 empirically selected covariates that were identified within Medicare files containing medication claims, inpatient and outpatient diagnostic codes and procedural codes. When selecting the 200 empirically selected covariates, we first identified the top 200 most prevalent codes within each data dimension (codes with a prevalence greater than 0.5 were subtracted from 1). We then selected the top 200 codes based on the strength of their univariate association (odds ratio) with the outcome. The estimated DRSs were implemented through stratification and again through 5-1 digit matching.[90]

To evaluate the validity of the fitted DRS model, we created a pseudo treatment group by sampling new-warfarin users within the original study cohort (i.e., index date after October 19, 2010). Sampling was done with replacement and with weights proportional to the odds of the PS. We created a pseudo control group by sampling with replacement from the same population, but with weights equal to one. We then evaluated the validity of the historically fitted DRS model by observing how well stratifying on the estimated scores controlled for confounding within the pseudo population. For comparison, we also evaluated the predictive performance of the estimated DRSs by assessing the calibration

(Hosmer-Lemeshow goodness of fit test) and discrimination (c-statistic) of the predicted values.

4.7 Empirical Results

We present results for the empirical study in Figure 4.3 and Tables 4.2 and 4.3. Table 4.2 shows the distribution of 37 a priori selected covariates across treatment groups. New-users of dabigatran were generally healthier with fewer comorbidities and greater use of the healthcare system than new-users of warfarin (Table 4.2). Similar patterns of initiation have been found in other studies.[81] Figure 4.3 shows similar PS distributions between the sampled pseudo population and original study cohort. Assuming the PS model is a close approximation to the true PS function, then the distribution of measured covariates across treatment groups and degree of confounding should also be similar across the pseudo and original study cohorts.

In Table 4.3 we present the unadjusted hazard ratio, as well as hazard ratios after DRS and PS matching. In Table 4.3 we also present measures for evaluating the validity of the historically estimated DRS as well as the PS model. Both PS and DRS matching resulted in similar effect estimates with hazard ratios of 0.88 (0.81, 0.95)and 0.87 (0.80, 0.94) respectively. The fitted PS model resulted in good predictive performance in terms of discrimination and calibration with a c-statistic of 0.73 and hosmer-lemeshow p-value of 0.52 (Table 4.3). The fitted DRS also resulted in good discrimination with a c-statistic of 0.78, but poor calibration in terms of the hosmer-lemeshow goodness-of-fit test with a p-value of <0.01 (Table 4.3). After matching on the PS, treatment groups were approximately balanced

on measured covariates with an ASAMD of <0.01, while DRS matching resulted in a pseudo bias of approximately 0.01 (Table 4).

4.8 Discussion

In this study, we used simulations and an empirical example to compare various measures for evaluating DRS models in their ability to reduce bias in effect estimates. We considered three measures of predictive performance including the c-statistic, the p-value from the Hosmer-Lemeshow goodness of fit test, and the MSE of the predicted values. We also evaluated the performance of the "dry run" method proposed by Hansen[73] where the fitted DRS is evaluated within a "pseudo population" of individuals who are sampled from the control population to create pseudo treatment and pseudo control groups that are representative of the original study cohort. In simulations, the calculated pseudo bias from the "dry run" had the strongest correlation with the bias in the treatment effect estimate when the functional form of the PS was either correctly specified or a close approximation to the true PS model. When there was moderate to strong misspecification in the PS, there was little to no correlation between the calculated pseudo bias from the dry run with the bias in the effect estimate.

Among the measures that evaluated the predictive performance of the fitted DRS models, the c-statistic and mse had the strongest correlation with bias in the effect estimate when comparing various models fit to the same data. In practice, however, researchers will often fit a single model (e.g., high-dimensional PS or DRS) and want to make a decision if the model is adequate in terms of confounding control. The c-statistic and MSE do not provide the best information for researchers when making a decision on a single DRS

model. In this study, we found that the c-statistic and MSE showed little to no correlation with bias in the effect estimate after combining measures across various data generating models.

Overall, we found that measures for evaluating the predictive performance of DRS models did not always correspond well with reduced bias in the estimated treatment effect. Previous studies have reported similar findings when evaluating PS models.[33, 34, 91] To what extent measures of predictive performance should be used when evaluating summary scores remains uncertain. Measures of predictive performance do not directly evaluate the ability of summary scores to control confounding. In contrast, measures of covariate balance across treatment groups when fitting the PS and the calculated pseudo bias within a "dry run" analysis when fitting the DRS can provide more direct measures for assessing the ability of the fitted models to control confounding.

Creating a pseudo population that is representative of the original study cohort requires accurate estimation of the PS. In this case, one could simply use the PS for confounding control. The DRS, however, has some desirable qualities that can be beneficial to researchers even when a correctly specified PS is available. DRSs provide a natural measure to evaluate treatment effect heterogeneity. When making treatment decisions, clinicians are almost always concerned about how the effect of a treatment varies over various patient profiles affecting the risk for the outcome (e.g., 10 year risk for cardiovascular disease based on the Framingham risk score). Although the PS allows us to detect and account for treatment effect heterogeneity, it does not provide the best information for health care providers in determining what subgroups of the patient

population are most likely to benefit from a given treatment regime. Further, matching or stratifying on the PS can be more restrictive then matching or stratifying on the DRS. The DRS can potentially allow for a greater number of individuals to be compared across treatment groups than the PS. This can be beneficial when there is strong separation in PS distributions (e.g., strong channeling with newly introduced treatments).

As with any simulation and empirical example, results are limited to the scenarios assessed. More research is needed to evaluate the performance of the discussed methods over a wide range of settings specific to large database research. Further, the optimal strategy for sampling from the control population when performing a "dry run" analysis remains unclear. More research is needed to evaluate various sampling strategies when creating pseudo treatment and pseudo control groups. Finally, neither measures of covariate balance for the PS or a "dry run" analysis for the DRS provide information on bias caused by unmeasured confounding or proper variable selection. We therefore stress the importance of using subject matter expertise to gain an understanding of the underlying causal structure before performing PS or DRS analyses.[83]

We conclude that accurately modeling the DRS within the study cohort, or within a historical set of controls presents unique challenges that are not shared by the PS. Measures of predictive performance do not necessarily identify the ability of a DRS model to control confounding. If the PS can be accurately modeled, evaluating the ability of the DRS model to control confounding within a "dry run" analysis can provide insight into the validity of fitted DRS models.

4.9 Tables and Figures

Table 4.1. Simulation Scenarios

Scenario	Distribution of Covs		Direction of Confounding	Baseline Prev of T and Y	Type of covariates		Sample size (n)
	Binomial(0.5)	Normal(0,1)			Confounder	Risk factor	
A	$X_1, X_3, X_5, X_6, X_8, X_9$	X_2, X_4, X_7, X_{10}	Both directions	50%	$X_1 - X_{10}$	-----	10,000
B	$X_1, X_3, X_5, X_6, X_8, X_9$	X_2, X_4, X_7, X_{10}	Both directions	10%	$X_1 - X_{10}$	-----	10,000
C	-------	$X_1 - X_{10}$	Both directions	50%	$X_1 - X_{10}$	-----	10,000
D	$X_1, X_3, X_5, X_6, X_8, X_9$	X_2, X_4, X_7, X_{10}	Both directions	50%	$X_1 - X_{10}$	-----	1,000
E	$X_1, X_3, X_5, X_6, X_8, X_9$	X_2, X_4, X_7, X_{10}	Both directions	50%	$X_1 - X_6$	$X_7 - X_{10}$	10,000
F	$X_1, X_3, X_5, X_6, X_8, X_9$	X_2, X_4, X_7, X_{10}	Same direction	50%	$X_1 - X_{10}$	-----	10,000

Table 4.2: Baseline covariates during 1 year washout period

	Warfarin (N=56,260)	Dabigatran 150mg (N=11,407)
Demographics:		
Age	78.91	76.76
Race (1 white, 0 other) (%)	89.2	91.72
Sex (% female)	42.17	48.95
% Diagnoses		
Cardiovascular:		
Chest pain	38.41	35.05
Heart disease	74.56	66.62
Heart failure	30.74	19.23
Hypertension	65.08	63.30
Hyperlipidemia	35.21	41.09
Myocardial Infarction	3.49	1.89
Cerebrovascular disease	21.29	17.38
Stroke		
Ischemic	6.09	4.31
Hemorrhagic	0.34	0.16
TIA	6.9	6.34
VTE	10.36	1.67
Diabetes	35.09	30.02
Kidney disease	12.58	4.74
Renal failure	16.09	5.75
Bleeding	1.88	0.68
Anemia	15.63	9.95
Baseline Meds: (%)		
Anti-depressants	28.27	22.89
Antihypertensives:		
ACE/ARB	52.22	50.23
Loop diuretics	40.91	28.70
Nonloop diuretics	52.55	41.97
Hypolipidemic drugs:		
Statins	49.40	52.45
Fibrate	5.02	4.98
Rate Control Therapy:		
Beta blockers	70.83	71.99
CCB	43.97	41.80
Glycoside	18.49	17.10
Rhythm Control Therapy	19.10	23.21
Healthcare Use (average #):		
# ECG claims	3.74	3.80
# PSA claims	0.36	0.49
# of fecal occult blood tests	0.12	0.13
# colonoscopies	0.14	0.14
# flu shot claims	0.76	0.79
# of lipid assessments	1.52	1.72
# of mammography claims	0.25	0.29
# of PapSmear claims	0.05	0.07

Table 4.3. Empirical results comparing dabigatran vs warfarin in the Medicare population between October 19, 2010 through December 31, 2013 (n=67,667)

# covs[a]	Method	HR (95% CI)	Pseudo bias[a]	ASAMD[b]	c-statistic	Hosmer lemeshow test[c]
237						
	Unadjusted	0.48 (0.46, 0.50)	0.45	0.12	-----	-----
	PS matching	0.88 (0.81, 0.95)	-----	0.01	0.73	p=0.52
	DRS matching	0.87 (0.80, 0.94)	0.01	-----	0.78	p<0.01

[a] PS and DRS models included 200 empirically selected covariates and 37 covariates selected a priori.

[b] average standardized absolute mean difference of covariates across treatment groups

[c] p-value from Hosmer-Lemeshow goodness of fit test

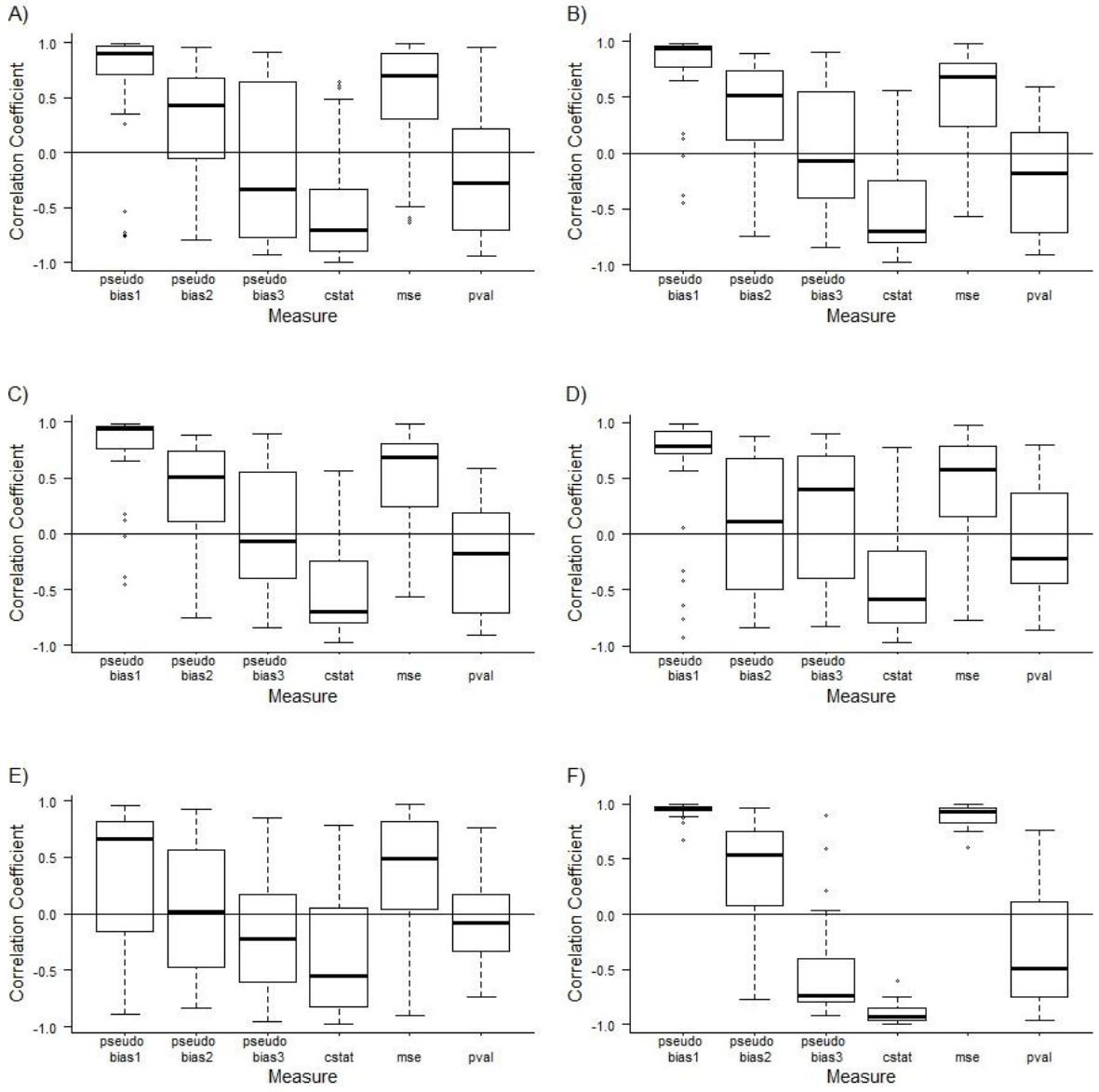

Figure 4.1. Box-plots of the correlation coefficients between each measure and absolute bias in the effect estimate for all parameter combinations. Each box-plot contains the correlation coefficients for the given measure across all parameter combinations. Plot a) shows the box plots for the basic scenario; plot b) the scenario containing low treatment and outcome prevalence, plot c) the scenario containing all continuous variables; plot d) small sample size; plot e) both confounders and risk factors in DRS; and plot f) confounding only in one direction.

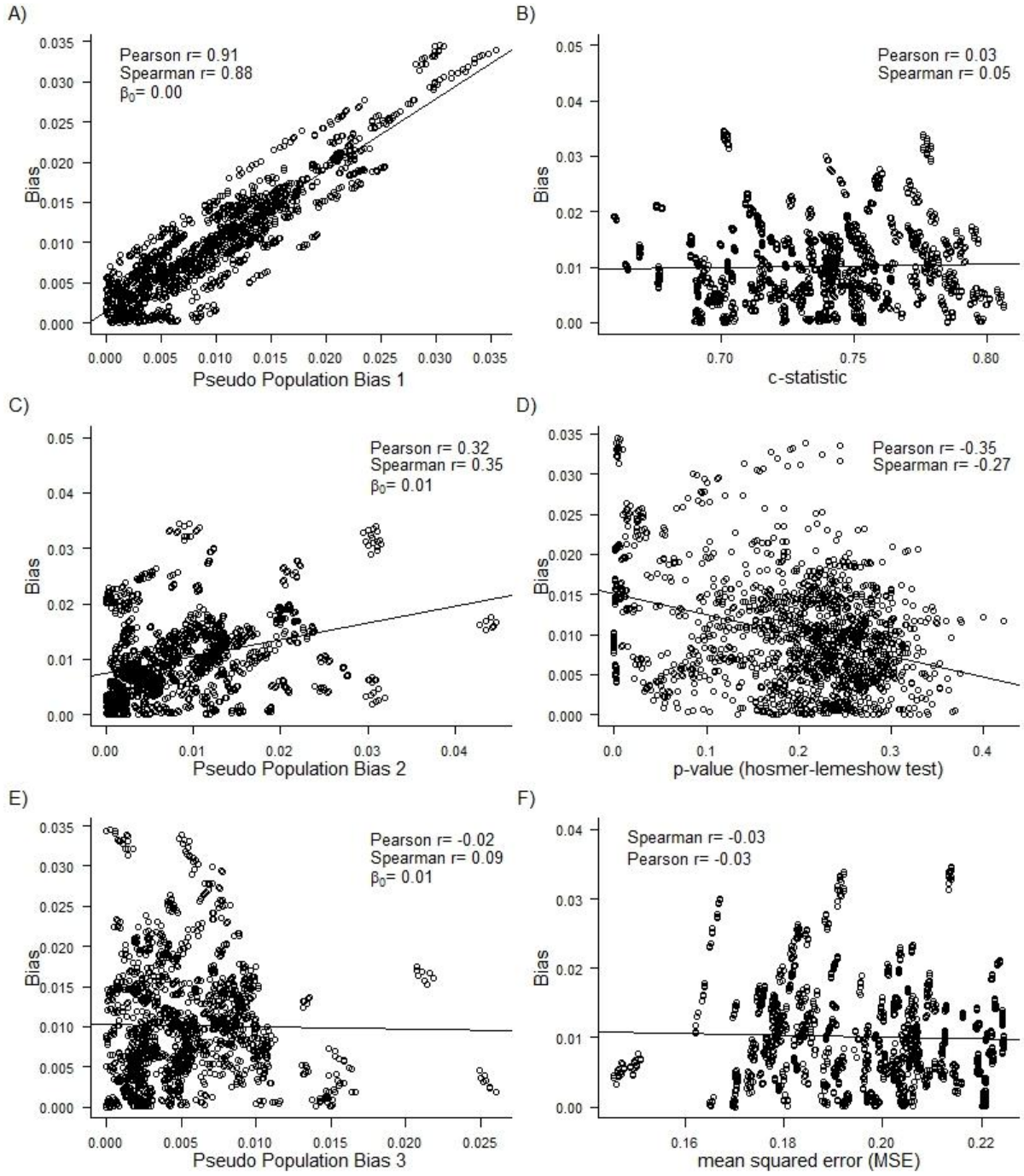

Figure 4.2. Measures for evaluating DRS models plotted against the absolute bias in the effect estimate for all parameter combinations in the basic scenario. Pseudo bias 1 in plot a) is the absolute bias within the pseudo population when the PS is mildly misspecified. Pseudo bias 2 in plot c) represents the bias in the pseudo population when the PS model is moderately misspecified, and pseudo bias 3 in plot e) when the PS model is strongly misspecified.

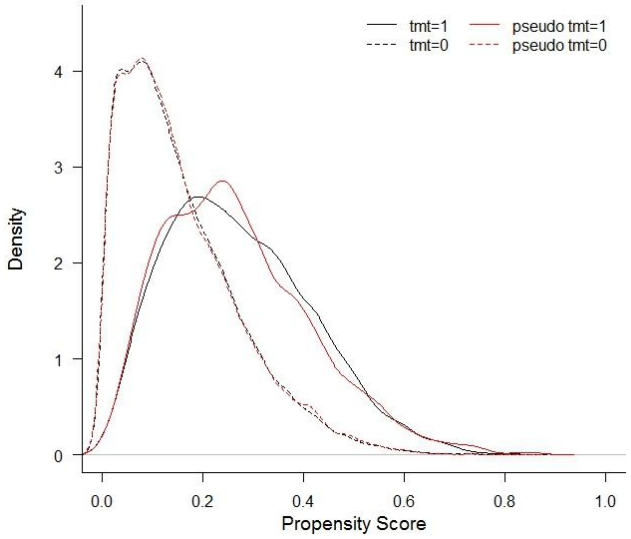

Figure 4.3. Propensity score distributions from original study cohort of dabigatran and warfarin new-users plotted against the pseudo propensity score distributions from the pseudo population of warfarin new-users.

CHAPTER 5

CONCLUSIONS & PUBLIC HEALTH SIGNIFICANCE

5.1 Summary of Specific Aims

Health care providers need rapid evaluation of newly marketed drugs to make timely decisions and optimize patient care. Large head-to-head clinical trials remain the gold standard when evaluating the efficacy of a given treatment, but can have limited generalizability and require long periods of time to complete. Electronic health care databases can provide timely information on patient populations who take newly introduced treatments in real time. Analyzing these data, however, present methodological challenges. Strong channeling and selective prescribing can lead to confounding by indication, requiring statistical methods and appropriate study design when evaluating these data.

Propensity scores have become widely used for controlling large numbers of confounding variables in large database research. It has been hypothesized, however, that the performance of PS methods may be limited when evaluating newly introduced treatments. A historically estimated disease risk score has been proposed as an alternative method for controlling large numbers of covariates in these settings. However, accurately modeling the disease risk score presents many challenges and it remains unclear if the use of disease risk scores in these settings is advantageous.

5.1.1 Summary of Aim 1

Using Medicare data, we examined the performance of using a historically estimated disease risk score when evaluating the comparative effectiveness of newly marketed drugs. We focused on evaluating the comparative effectiveness of the new oral-anticoagulant dabigatran with warfarin in preventing combined ischemic stroke and all-cause mortality. Currently, there is limited information available on the net beneficial gains that new oral anticoagulant medications have on cardiovascular events compared to warfarin in real-world practice.

Due to the very recent approval of these new oral anticoagulants, their evaluation in non-experimental settings is difficult, in part due to the limited data available to control for large sets of confounders. It has been hypothesized that out-of-sample estimation methods for disease risk can be advantageous because these methods will allow for the control of a large number of risk factors at the start of drug introduction, potentially allowing researchers to evaluate the comparative effectiveness of these newer medications at earlier periods than previous methods.

When comparing dabigatran with warfarin, we found that dabigatran new-users tended to be younger and healthier than new-users of warfarin. After controlling for a high-dimensional set of covariates, effect estimates were more consistent with clinical trials. Controlling for a high-dimensional set of baseline covariates can improve confounding control, but can also create separation in the PS distributions across treatment groups, limiting the number of exchangeable individuals within the study cohort. When PS distributions are separated, we found that the DRS can allow researchers to compare the

treatment effect within a larger proportion of the population, potentially improving precision and the accuracy of the treatment effect when the parameter of interest is the average treatment effect in the full treated population. In this study, we found that this benefit of the DRS is most pronounced with smaller sample sizes. Finally, while it has been hypothesized that the DRS can be more stable over time potentially simplifying its estimation compared to the PS for newly marketed drugs, in our example we found modeling the DRS to be more challenging than modelling the PS. In general, modeling the DRS presents more challenges than modelling the PS, even in settings involving new treatments. Reporting results from PS analyses in addition to analyses using a historically estimated DRS can be beneficial in comparative effectiveness research of new treatments.

5.1.2 Summary of Aim 2

Accurately modeling the DRS, either within historical set of controls or the original study cohort, presents challenges that are not shared when modeling the PS. These difficulties highlight the importance of evaluating the validity of fitted DRS models. Researchers have primarily evaluated risk models by assessing their predictive performance in terms of discrimination (e.g., c-statistic) and calibration (e.g., goodness of fit tests). In this study, we found that measures for evaluating the predictive performance of DRS models did not always correspond well with reduced bias in the estimated treatment effect. In contrast, measures of covariate balance across treatment groups when fitting the PS and the calculated pseudo bias within a "dry run" analysis when fitting the DRS can provide more direct measures for assessing the ability of summary scores to control confounding.

Creating a pseudo population that is representative of the original study cohort requires accurate estimation of the PS. In this case, one could simply use the PS for confounding control. The DRS, however, has some desirable qualities that can be beneficial to researchers even when a correctly specified PS is available. DRSs provide a natural measure to evaluate treatment effect heterogeneity and can allow for a greater number of individuals to be compared across treatment groups than the PS. This can be beneficial when there is strong separation in PS distributions (e.g., strong channeling with newly introduced treatments). In conclusion, the DRS can be beneficial when evaluating newly introduced treatments. Finding more accurate ways to evaluate the validity of fitted DRS models can improve the quality of the estimation of disease risk scores. Hansen's proposed method of evaluating the fitted DRS within a "dry run" analysis is promising, but more research is needed over a range of settings specific to large database research.

REFERENCES

1. Rosenbaum PR R, DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.

2. 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(5):437-47.

3. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. Pharmacoepidemiology and drug safety. 2012;21 Suppl 2:138-47.

4. Mack CD, Glynn RJ, Brookhart MA, et al. Calendar time-specific propensity scores and comparative effectiveness research for stage III colon cancer chemotherapy. Pharmacoepidemiology and drug safety. 2013;22(8):810-818.

5. Hansen BB. The prognostic analogue of the propensity score. Biometrika. 2008;95(2):481-8. doi:10.1093/biomet/asn004

6. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. Statistical methods in medical research. 2009;18(1):67-80.

7. Arbogast PG, Ray WA. Performance of Disease Risk Scores, Propensity Scores, and Traditional Multivariable Outcome Regression in the Presence of Multiple Confounders. American journal of epidemiology. 2011;174(5):613-20.

8. Singh M, Adigopula S, Patel P, Kiran K, Khosla S. Recent advances in oral anticoagulation for atrial fibrillation. Therapeutic advances in cardiovascular disease. 2010;4(6):395-407.

9. Cabral KP, Ansell J, Hylek EM. Future directions of stroke prevention in atrial fibrillation: the potential impact of novel anticoagulants and stroke risk stratification. Journal of thrombosis and haemostasis : JTH. 2011;9(3):441-9.

10. Adam SS, McDuffie JR, Ortel TL, Williams JW, Jr. Comparative effectiveness of warfarin and new oral anticoagulants for the management of atrial fibrillation and venous thromboembolism: a systematic review. Annals of internal medicine. 2012;157(11):796-807.

11. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. American journal of epidemiology. 2003;158(9):915-20.

12. Ray WA. Population-based studies of adverse drug effects. The New England journal of medicine. 2003;349(17):1592-4.

13. Rubin DB. Estimating causal effects from large data sets using propensity scores. Annals of internal medicine. 1997;127(8 Pt 2):757-63.

14. R L. Beyond clinical trials: The importance of large databases in evaluating differences in the effectiveness of bisphosphonate therapy in posmenopausal osteoporosis. Bone. 2007;40:S32-S5.

15. Ray WA. Improving automated database studies. Epidemiology. 2011;22(3):302-4.

16. Sentinel Initiative. http://www.fda.gov/safety/FDAsSentinelInitiative/ucm2007250.htm. Last accessed March 4, 2013.

17. Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. Basic & clinical pharmacology & toxicology. 2006;98(3):253-9.

18. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688-701.

19. Rubin D. Causal inference using potential outcomes. J Amer Statist Assoc. 2005;100(469):322-31.

20. Sekhon J. The Neyman_Rubin model of causal inference and estimation via matching methods. The Oxford Handbook of Political Methodology. 2007.

21. Heckman J, Ichimura, H, Todd, P. Matching as an econometric evaluation estimator. Review of Economic Studies. 1998;65:261-94.

22. PR JMaR. Invited Commentary: Propensity Scores. American journal of epidemiology. 1999;150(4):327-33.

23. Tadrous M, Gagne JJ, Sturmer T, Cadarette SM. Disease risk score as a confounder summary method: systematic review and recommendations. Pharmacoepidemiology and drug safety. 2013;22(2):122-9.

24. 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(9):891-8.

25. CC P. A Method of Matching Groups for Experiment with No Loss of Population. The Journal of Educational Research. 1941;34(8):606-12.

26. WA B. A Technique for Studying the Effects of a Television Broadcast. Journal of the Royal Statistical Society. Series C. 1956;5(3):195-202.

27. Miettinen OS. Stratification by a multivariate confounder score. American journal of epidemiology. 1976;104(6):609-20.

28. 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(1):104-6.

29. 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(4):317-24.

30. Leacy FP, Stuart EA. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. Statistics in medicine. 2014;33(20): 3488-508.

31. Cadarette SM, Gagne JJ, Solomon DH, Katz JN, Sturmer T. Confounder summary scores when comparing the effects of multiple drug exposures. Pharmacoepidemiology and drug safety. 2010;19(1):2-9.

32. Wyss R, Lunt M, Brookhart MA, Glynn RJ, Stürmer T. Reducing bias amplification in the presence of unmeasured confounding through out-of-sample estimation strategies for the disease risk score. J. Causal Infer. 2014;2(2): 131-146.

33. Westreich D, Cole SR, Funk MJ, Brookhart MA, Sturmer T. The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiology and drug safety. 2011;20(3):317-20.

34. Wyss R, Ellis AR, Brookhart MA, et al. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. American journal of epidemiology. 2014;180(6):645-55.

35. Conen D CC, Glynn RJ, Tedrow UB, Everett BM, Buring JE, Albert CM. Risk of death and cardiovascular events in initially healthy women with new-onset atrial fibrillation. JAMA. 2011;305(20):2080-7.

36. Khoo CW, Lip GY. Burden of atrial fibrillation. Current medical research and opinion. 2009;25(5):1261-3.

37. Stewart S, Hart CL, Hole DJ, McMurray JJ. A population-based study of the long-term risks associated with atrial fibrillation: 20-year follow-up of the Renfrew/Paisley study. The American journal of medicine. 2002;113(5):359-64.

38. Wang TJ, Larson MG, Levy D, et al. Temporal relations of atrial fibrillation and congestive heart failure and their joint influence on mortality: the Framingham Heart Study. Circulation. 2003;107(23):2920-5.

39. Lloyd-Jones DM, Wang TJ, Leip EP, et al. Lifetime risk for development of atrial fibrillation: the Framingham Heart Study. Circulation. 2004;110(9):1042-6.

40. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation: a major contributor to stroke in the elderly. The Framingham Study. Archives of internal medicine. 1987;147(9):1561-4.

41. Patel MR, Mahaffey KW, Garg J, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. The New England journal of medicine. 2011;365(10):883-91.

42. Miyasaka Y, Barnes ME, Gersh BJ, et al. Secular trends in incidence of atrial fibrillation in Olmsted County, Minnesota, 1980 to 2000, and implications on the projections for future prevalence. Circulation. 2006;114(2):119-25.

43. Mega JL. A new era for anticoagulation in atrial fibrillation. The New England journal of medicine. 2011;365(11):1052-4.

44. Jones M, McEwan P, Morgan CL, Peters JR, Goodfellow J, Currie CJ. Evaluation of the pattern of treatment, level of anticoagulation control, and outcome of treatment with warfarin in patients with non-valvar atrial fibrillation: a record linkage study in a large British population. Heart. 2005;91(4):472-7.

45. Choudhry NK, Anderson GM, Laupacis A, Ross-Degnan D, Normand SL, Soumerai SB. Impact of adverse events on prescribing warfarin in patients with atrial fibrillation: matched pair analysis. BMJ. 2006;332(7534):141-5.

46. Choudhry NK, Soumerai SB, Normand SL, Ross-Degnan D, Laupacis A, Anderson GM. Warfarin prescribing in atrial fibrillation: the impact of physician, patient, and hospital characteristics. The American journal of medicine. 2006;119(7):607-15.

47. Cohen N, Almoznino-Sarafian D, Alon I, et al. Warfarin for stroke prevention still underused in atrial fibrillation: patterns of omission. Stroke; a journal of cerebral circulation. 2000;31(6):1217-1222.

48. Fang MC, Stafford RS, Ruskin JN, Singer DE. National trends in antiarrhythmic and antithrombotic medication use in atrial fibrillation. Archives of internal medicine. 2004;164(1):55-60.

49. Connolly SJ, Ezekowitz MD, Yusuf S, et al. Dabigatran versus warfarin in patients with atrial fibrillation. The New England journal of medicine. 2009;361(12):1139-51.

50. Granger CB, Alexander JH, McMurray JJ, et al. Apixaban versus warfarin in patients with atrial fibrillation. The New England journal of medicine. 2011;365(11):981-92.

51. Alexander GC, Sehgal NL, Moloney RM, Stafford RS. National trends in treatment of type 2 diabetes mellitus, 1994-2007. Archives of internal medicine. 2008;168(19):2088-94.

52. Ansell J. New oral anticoagulants should not be used as first-line agents to prevent thromboembolism in patients with atrial fibrillation. Circulation. 2012;125(1):165-70; discussion 70.

53. De Caterina R, Husted S, Wallentin L, et al. New oral anticoagulants in atrial fibrillation and acute coronary syndromes: ESC Working Group on Thrombosis-Task Force on Anticoagulants in Heart Disease position paper. Journal of the American College of Cardiology. 2012;59(16):1413-25.

54. Kaluski E, Maher J, Gerula CM. New oral anticoagulants: good but not good enough! Journal of the American College of Cardiology. 2012;60(15):1434; author reply -5.

55. Winkelmayer WC, Liu J, Setoguchi S, Choudhry NK. Effectiveness and safety of warfarin initiation in older hemodialysis patients with incident atrial fibrillation. Clinical journal of the American Society of Nephrology : CJASN. 2011;6(11):2662-8.

56. Hess PL, Greiner MA, Fonarow GC, et al. Outcomes associated with warfarin use in older patients with heart failure and atrial fibrillation and a cardiovascular implantable electronic device: findings from the ADHERE registry linked to Medicare claims. Clinical cardiology. 2012;35(11):649-657.

57. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Medical care. 2010;48(6 Suppl):S114-20.

58. Schneeweiss S, Patrick AR, Sturmer T, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. Medical care. 2007;45(10 Supl 2):S131-42.

59. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Medical care. 2005;43(5):480-5.

60. Kokotailo RA, Hill MD. Coding of stroke and stroke risk factors using international classification of diseases, revisions 9 and 10. Stroke; a journal of cerebral circulation. 2005;36(8):1776-81.

61. 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(4):512-22.

62. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. American journal of epidemiology. 2006;163(12):1149-56.

63. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. American journal of epidemiology. 2011;174(11):1213-22.

64. Redelmeier DA, Tan SH, Booth GL. The treatment of unrelated disorders in patients with chronic medical diseases. The New England journal of medicine. 1998;338(21):1516-20.

65. Welch HG, Albertsen PC, Nease RF, Bubolz TA, Wasson JH. Estimating treatment benefits for the elderly: the effect of competing risks. Annals of internal medicine. 1996;124(6):577-84.

66. Walker AM. Confounding by indication. Epidemiology. 1996;7(4):335-6.

67. Glynn RJ, Knight EL, Levin R, Avorn J. Paradoxical relations of drug treatment with mortality in older persons. Epidemiology. 2001;12(6):682-9.

68. Brookhart MA, Patrick AR, Dormuth C, et al. Adherence to lipid-lowering therapy and the use of preventive health services: an investigation of the healthy user effect. American journal of epidemiology. 2007;166(3):348-54.

69. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological methods. 2004;9(4):403-25.

70. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Statistics in medicine. 2010;29(3):337-46.

71. Hosmer DWL, S. A goodness-of-fit test for the multiple logistic regression model. Communications in Statistics 1980;A10:1043-69.

72. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford, UK: Oxford University Press2003.

73. Hansen BB. Bias reduction in observational studies via prognosis scores. : Statistics Department, University of Michigan, Ann Arbot, Michigan 2006.

74. Chen XD, A.P., Liu, J.S. Weighted finite population sampling to maximize entropy. Biometrika. 1994;81(3):457-69.

75. 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 Supl 2):S143-8.

76. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. American journal of epidemiology. 2011;173(12):1404-13.

77. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974;66:688-701.

78. Franklin JM, Rassen JA, Bartels DB, Schneeweiss S. Prospective cohort studies of newly marketed medications: using covariate data to inform the design of large-scale studies. Epidemiology. 2014;25(1):126-33.

79. Schneeweiss S, Gagne JJ, Glynn RJ, Ruhl M, Rassen JA. Assessing the comparative effectiveness of newly marketed medications: methodological challenges and implications for drug development. Clinical pharmacology and therapeutics. 2011;90(6):777-90.

80. Gagne JJ, Bykov K, Willke RJ, Kahler KH, Subedi P, Schneeweiss S. Treatment dynamics of newly marketed drugs and implications for comparative effectiveness research. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2013;16(6):1054-62.

81. Desai NR, Krumme AA, Schneeweiss S, et al. Patterns of Initiation of Oral Anticoagulants in Patients with Atrial Fibrillation - Quality and Cost Implications. The American journal of medicine. 2014.

82. Graham DJ, Reichman ME, Wernecke M, et al. Cardiovascular, Bleeding, and Mortality Risks in Elderly Medicare Patients Treated with Dabigatran or Warfarin for Non-Valvular Atrial Fibrillation. Circulation. 2014.

83. Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology. 2001;12(3):313-20.

84. Kumamaru HG, JJ; Glynn, RJ; Setoguchi, S; Schneeweiss, S. Dimension reduction and shrinkage methods for improving high dimensional disease risk score estimation in a historical cohort. Pharmacoepidemiology and drug safety. 2014;23(s1):267.

85. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. Statistics in medicine. 2014;33(10):1685-99.

86. Ali MS, Groenwold RH, Pestman WR, et al. Propensity score balance measures in pharmacoepidemiology: a simulation study. Pharmacoepidemiology and drug safety. 2014;23(8):802-11.

87. Rubin DB. On principles for modeling propensity scores in medical research. Pharmacoepidemiology and drug safety. 2004;13(12):855-7.

88. Imai KR, M. Covariate balancing propensity score. J R Stat Soc B. 2014;76(1):243-63.

89. Hernan MA, Robins JM. Estimating causal effects from epidemiological data. Journal of epidemiology and community health. 2006;60(7):578-86.

90. Parsons LS. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. SUGI 26 Proceedings: Cary, NC: SAS Institute2001.

91. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. Pharmacoepidemiology and drug safety. 2005;14(4):227-38.