Application of Targeted Maximum Likelihood Estimation using Varying Lookback Windows in Pharmacoepidemiology

RUBIYA AKTER

Department of Epidemiology, Biostatistics and Occupational Health

McGill University, Montreal

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Science in Biostatistics

©RUBIYA AKTER

December 12, 2023

Abstract

In observational research, confounding control is critical to appropriate inference. In the field of pharmacoepidemiology, determining the appropriate lookback period is essential as it dictates how far back we examine potential confounding factors from the time of exposure. This study aims to compare treatment effect estimates using two different approaches: targeted maximum likelihood estimation (TMLE) and propensity score method under inverse probability weighting approach (PS-IPW), considering varying lookback periods (short term and long term). To conduct the comparison, simulation settings were established, considering eight different lookback periods (1,3,6 months, and 1,2,5,7,9 years), with 10 years as the ideal reference lookback period. We applied two different approaches: PS-IPW and TMLE-logistic, both within the logistic regression framework. Additionally, we included TMLE using SuperLearner (TMLE-SL) as part of our simulation. To assess the effect of lookback on propensity score models, propensity score quantile estimates were computed, revealing that longer lookback periods exhibited less bias compared to shorter ones. Subsequently, the average treatment effect (ATE) along with standard error was estimated using PS-IPW, TMLE-logistic, and TMLE-SL. It was observed that TMLE-logistic and TMLE-SL produced lower standard errors than the PS-IPW approach across the varying lookback periods. Finally, the study applied both methods to the CPRD (Clinical Practice Research Datalink) database to evaluate how TMLE and PS-IPW perform in real-life scenarios for each of the specified lookback periods. In conclusion, this research contributes valuable insights into the impact of lookback periods on treatment effect estimates, highlighting the advantages of using TMLE approaches over PS-IPW in certain lookback scenarios. Furthermore, the application of these methods to real-world data from CPRD aids in understanding their performance in practical healthcare research settings.

Abrégé

En recherche observationnelle, le contrôle des facteurs de confusion est essentiel pour obtenir des inférences appropriées. Dans le domaine de la pharmacoépidémiologie, la détermination de la période de référence appropriée est essentielle, car elle détermine à quel point nous examinons les facteurs de confusion potentiels à partir du moment de l'exposition. Cette étude vise à comparer les estimations des effets du traitement en utilisant deux approches différentes : l'estimation du maximum de vraisemblance ciblée (TMLE) et la méthode du score de propension avec approche de pondération par probabilité inverse (PS-IPW), en tenant compte de différentes périodes de référence (court terme et long terme). Pour mener la comparaison, des paramètres de simulation ont été établis en considérant huit périodes de référence différentes (1, 3, 6 mois, et 1, 2, 5, 7, 9 ans), avec une période de référence idéale de 10 ans. Nous avons appliqué deux approches différentes : PS-IPW et TMLE-logistique, toutes deux dans le cadre de la régression logistique. De plus, nous avons inclus TMLE utilisant SuperLearner (TMLE-SL) dans notre simulation. Pour évaluer l'effet de la période de référence sur les modèles de score de propension, des estimations des quantiles du score de propension ont été calculées, révélant que les périodes de référence plus longues présentaient moins de biais par rapport aux périodes plus courtes. Ensuite, l'effet moyen du traitement (ATE) ainsi que l'erreur standard ont été estimés en utilisant PS-IPW, TMLE-logistique et TMLE-SL. Il a été observé que TMLE-logistique et TMLE-SL produisaient des erreurs standard plus faibles que l'approche PS-IPW sur les différentes périodes de référence. Enfin, l'étude a appliqué ces deux méthodes à la base de données CPRD (Clinical Practice Research Datalink) pour évaluer comment TMLE et PS-IPW se comportent dans des scénarios réels pour chacune des périodes de référence spécifiées. En conclusion, cette recherche apporte des informations précieuses sur l'impact des périodes de référence sur

les estimations des effets du traitement, mettant en évidence les avantages de l'utilisation des approches TMLE par rapport à PS-IPW dans certains scénarios de période de référence. De plus, l'application de ces méthodes aux données du monde réel provenant de la base de données CPRD contribue à comprendre leur performance dans des contextes pratiques de recherche en santé.

Dedication

*To my daughter **ARIBA***

Acknowledgements

Above all else, I wish to convey my profound appreciation to my supervisor Dr. Robert Platt, for being an unwavering source of support throughout my graduate study. Despite his demanding schedule, I was fortunate to receive his comprehensive guidance and invaluable advice. Without his support and mentorship, the completion of this thesis would not have been possible.

My heartfelt thanks go to my graduate co-supervisor, Dr. Samy Suissa, and supervisor Dr. Robert Platt, as well as Lady Davis Institute (LDI), for providing financial support during my graduate studies. Their generous assistance allowed me to focus on my studies and research without financial burdens.

I am also sincerely grateful to the Department of Epidemiology, Biostatistics, and Occupational Health (EBOH) for cultivating a friendly and supportive academic environment. I feel humbled and honored to have received the Graduate Excellence Award in September 2022 from this department. This recognition further motivates me to strive for excellence in my future endeavors.

I would like to express my gratitude to Professor Kristian B. Filion for facilitating my access to CPRD data. Additionally, I would like to thank Professor Enrico Ripamonti for his insightful suggestions during the simulation study of this thesis. My special thanks to Marisa Mancini for her exceptional administrative support during my time at the university and to the LDI IT team for their technical support.

Lastly, I want to extend my heartfelt appreciation to my family, especially my husband, for being a constant source of support throughout the last two years of my master's journey. His continuous encouragement helped me navigate through all the challenges I encountered, and I am truly grateful for his presence in my life.

Preface

I wrote this dissertation under the direction and guidance of Dr. Robert Platt. I have done all the simulations, data analysis and written all the chapters.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Pharmacoepidemiologic research is fundamentally interested in the estimation of the causal effect of medications on safety and efficacy outcomes. To estimate the causal effect of exposure on a disease outcome, the randomized controlled trial (RCT) is useful as random allocation of treatment removes confounding bias. However, the RCT design is mainly implemented in pre-market studies to adopt a new intervention and compare its effectiveness with the current drugs in the market [1]. Due to limited budget, low resources, and ethical considerations, the RCT design is not always feasible once the drug is on the market [2]. Further, the result from an RCT lacks generalizability since it is mostly limited to a selected homogeneous population, and the sample size is typically small relative to the population using the medication. So, this design has less ability to detect uncommon but important harms. Also, in clinical decision-making, some interventions may never be subject to randomization. Thus, observational studies are sometimes the only option to get data on specific scientific questions and to help make decisions about limited health care budget allocation [3, 4, 5]. Moreover, after the approval of a new drug, further investigation tends to be continued to ensure drug safety in humans [6, 7, 8]. Therefore to identify previously unrecognized adverse effects as well as positive effects of drugs, observational research is adopted in post-marketing surveillance [9, 10, 11]. A crucial problem in observational studies emerges when the treatment groups are not directly comparable due to covariate imbalances (i.e., confounding). To minimize this confounding bias in observational studies, several methods have been developed in the literature [12]. Propensity score

methods are one of the most widely used [13, 14, 15, 16, 17]. The propensity score methods were first introduced in 1987 by Rosenbaum [18] and are commonly used to balance between treatment and comparison groups. The propensity score is the probability of treatment assignment conditioning on some observed baseline characteristics. To design and analyze an observational study, propensity score methods imitate some particular characteristics of a randomized control trial by providing a balancing score. In particular, the propensity score is a balancing score such that, when conditioning on the propensity score, the treated and untreated groups would have a similar covariate distribution on average. There are four well-known propensity score methods, i.e., matching on the propensity score, stratification on the propensity score, inverse probability treatment weighting using the propensity score, and covariate adjustment using the propensity score.

Observational healthcare data, especially administrative healthcare data, are often criticized for partial or incomplete information about patients' past history. In the presence of confounding, drawing appropriate inferences is critical. So, it is important to identify the lookback period to determine how far back from exposure one looks to assess potential confounders [19, 20, 21]. Typically researchers often decide on a lookback time window. A previous methodological study (Ripamonti et al., under review in Statistical Methods in Medical Research) has been done to explore how much history should be considered in propensity score estimation. The study showed that with longer lookback time windows, the estimates of the true propensity score distribution are less biased and more precise than with shorter lookback time [19].

However, the propensity score method has some limitations as it requires the treatment model to be correctly specified given confounders. To obtain an unbiased estimate of the average treatment effect (ATE), correct model specification is necessary [22, 23, 24]. To overcome this challenge, a semi-parametric double robust method called targeted maximum likelihood estimation (TMLE) has been introduced [25]. This method is robust to misspecification of either treatment or outcome model, making it appealing in confounding adjustment over other naive approaches. Besides, it allows the incorporation of flexible ma-

chine learning algorithms for choosing the correct model. But due to its limited familiarity, TMLE has not been extensively used in observational drug studies. To our knowledge, The estimation of average treatment effect (ATE) using TMLE in the context of varying lookback time lengths has not yet been explored. This study explores estimating the average treatment effect using TMLE under various lookback settings. Moreover, we will explore how the estimate of ATE can be affected using a data-driven approach by defining long term and short term lookback periods. Finally, we intend to compare propensity score methods and TMLE in various timestamped lookback data under a simulation study and real world settings.

Chapter 2

Literature Review

2.1 Observational Studies in Pharmacoepidemiology

The field of pharmacoepidemiology uses observational studies to examine drug safety and effectiveness [26, 9]. In this modern age of evidence-based medicine, clinical researchers require an extensive range of well-designed studies to provide an appropriate prescribing decision for patients. Observational studies have become necessary to provide evidence to assess drug safety and detect the advantages and disadvantages of approved medications [27]. Randomized studies allow us to directly compare the results from those assigned treatment to those from the control group. To ensure a fair comparison between comparable groups, we randomly allocate the subjects into one of the two treatment groups. In randomized controlled trials (RCT), the sample size is typically planned to be large enough to fairly compare two treatment groups on primary efficacy outcomes. However, clinical trials are run on a restricted study population determined by the pre-approved protocol. For example, testing a new intervention on pregnant women or children is usually not ethical. RCTs typically focus on a population that is easily managed and less heterogenous than the general population. Thus the result from an RCT may be locally valid but not admissible to check the effectiveness or safety of treatment in the general population. In that case, observational studies are the only way to collect information about a particular treatment or medication. Observational studies include cohort or case-control studies [28].

A key feature of the observational study is that it is designed so that the study does not dictate any intervention on the participants. Instead, it observes their drug usage and health status. The study selects a cohort of people with a specific disease of interest or a combination of characteristics and reports their health information at periodic intervals. Since no experimental drug is prescribed, ethical concerns are reduced in the observational study design. Thus the participants in each study cohort become more representative and reflect "real life" drug usage. The main drawback of such study designs is that the exposure or intervention is not randomized in two treatment groups, and therefore **confounding** by indication [29] becomes a problem. Confounding is a source of bias that corresponds to lack of comparability between treatment and exposure groups when estimating the causal effect [29]. Confounding bias may arise due to differences in any measured or unmeasured confounders between two comparison groups. Any observed exposure effect on the outcome may be due to these baseline differences. Various statistical tools have been introduced in the literature to address confounding in this observational study setting.

Due to improvements in observational study methods and advanced statistical techniques, researchers can investigate real life data, rare outcomes, and long term effects that were unexplored in pre-approval RCTs [30, 31]. In a review of empirical studies, a study suggested that systematic review or meta-analysis based on RCT and well-conducted observational studies produce similar effect estimates in many settings [32]. The discrepancies between these studies might happen due to study design or other causes, and in such cases, appropriate clinical reasoning and causal inference should be provided [32]. Another study of meta-analyses has shown that the RCT and observational studies have more similar results than have often been thought [33]. An important feature of observational research is that it can include millions of patients to study drug safety, which is unlikely to happen in randomized controlled trials because of limited size and follow up. The databases used in the observational study are typically from patient registries, electronic health records, routinely collected administrative data, primary patient-level data collection (prospective or retrospective), or population health surveys [34]. Information on medication from these data sources is usually more accurate than self-recorded information, particularly in the

situation when patients are too ill or have died already. A study suggested that observational studies may be less prone to heterogeneity and more representative of clinical practice because they have large populations [35]. Moreover, observational research can address research questions that are not suitable for randomized clinical trials. For example, observational studies can address safety outcomes, and there can be individuals whose administration is not consistent with treatment guidelines. Proper use of observational studies can investigate disease prevalence, incidence, association, causes, and outcomes. Observational studies can provide critical, descriptive data when little is known about the epidemiology of disease and information on long term drug efficacy and safety in a cost-effective way. Typically these studies include case reports and case series, ecological, cross-sectional, cohort, and case-control studies. In pharmacoepidemiology, many studies use cohort study design [36].

Since observational studies cannot control for bias and confounding by design, statistical tools must be used for confounding adjustment. As pharmacoepidemiologic research takes advantage of large databases because of the growing trend of recorded electronic data, observational studies use these retrospective databases that can engage longer observation periods and large populations [37]. Thus observational studies answer various research questions at less expense and without long delays and help decision-makers compare the efficacy of various medications [38].

2.2 Lookback Periods in Observational Studies

In observational studies, the researcher does not know the treatment assignment mechanism, as the treatment for an individual can be selected by itself or by a third party. The researcher starts with a comparison group that may not be equivalent to the treatment group in terms of background characteristics. To conduct a valid study, covariates related to the treatment assignment and potential outcomes should be identified and addressed by the researcher. For example, a researcher wants to evaluate a home tutoring program to improve students' science skills. The researcher has to consider all covariates related to math,

physics, and biology score. Some of these covariates may be observed or unobserved. The observed variables can be identified by the student's exam scores in corresponding subjects, prior experiences, and demographic characteristics. But there can be some unmeasured variables such as students' capability to learn, parents' attitude towards education, students' personality factors, etc. Failure to account for these variables may lead to biased or inconsistent effect estimates.

In pharmacoepidemiology, observational studies use administrative health care data routinely compiled by various health care providers and patient encounters [39, 40]. Often these data are criticized for the incompleteness of information on potential confounders [41]. Moreover, the past history of drug exposure is only partially available or missing because of the non-availability of the individuals under observation. This problem can lead to potential biases and thus questions can arise on the validity of the research being conducted. Rather than ignoring this data, researchers decide on a lookback period to identify individuals and to identify relevant confounders. Subjects can vary in terms of the duration of available baseline information. In this situation, the researchers have two choices, either use covariate information based on all available baseline data for each individual or use covariate information based on a fixed baseline window of time that is shared by all subjects [19]. Previous studies showed that assessing baseline confounders with fixed/short lookback time may cause biased or inconsistent effect estimates [42].

2.2.1 Review of the Choice of Optimal Lookback Length

Relevant information like past acute diseases and events might not be captured if the lookback window is not optimal. Previous research has investigated how this length is selected and showed how it affects results [43]. A Korean study using a health insurance database has shown the number of miss-classified incident cases decreases as the lookback length increases [44]. Another US study showed that a 3-year lookback could produce a less biased point estimate [45]. A study conducted by Osokogu [46], has shown that the length of various lookback periods might affect inference using propensity scores in electronic health care data. The study made a comparison among lookback periods of 1 week, 1 month, and

3 months and corresponding results showed that 3 months lookback periods might lead to the most comparable results to clinical trial results compared to other lookbacks. However, they suggested all available lookback is preferable over short fixed lookbacks. In addition, an Australian study using linked perinatal population data showed that although longer ascertainment (lookback) periods might improve the identification of chronic disease history, it did not change the C-statistic [47]. The study's findings indicated that information regarding comorbidity prior to hospital admissions had minimal impact on hemorrhage (an acute loss of blood from a damaged blood vessel) modeling. Instead, the study proposed determining an appropriate lookback period based on the specific objectives of the research [47]. These findings might help us to conclude that evaluating previous hospital history for the completeness of disease history is not always worthwhile. A US study was performed to quantify the effect of lookback periods on the misclassification of true new users of antibiotics and asthma medications [48]. The study defined lookback periods ranging from 30 days to 12 years and reported severe misclassification for 30 days to 2-year lookback. However, it suggested that studies with limited prescription history are likely to be unable to observe an adverse drug effect, and the choice of optimal lookback length should depend on the type of medication and the outcome that is being studied [48]. Another study aimed at quantifying the effects of different comorbidity ascertainment lookback lengths on modeling post-hospitalization mortality and readmission outcomes [49]. They defined lookback period 1 (1-365 days), lookback 2 (366 days- 2 years), lookback 3 (2-3 years), and final period (3-5 years) [49]. They reported that the predictive ability of regression models is affected by the length of the lookback. Lookback period 1 appeared to be appropriate for modeling posthospitalization mortality where longer lookback was preferable for readmission outcomes [49]. Several authors performed a systematic review of multiple comorbidity measures and suggested that a lookback period 1 year before the index date is preferable to improve mortality prediction [50]. However, another study was performed in Norway to examine the effect of lookback period length to identify acute MI (myocardial infarction) on incidence rates by performing a subgroup analysis between men and women [51]. The study concluded that the lookback period of 7 and 10 years are reliable to identifying the incident of acute myocardial infarction whereas the shorter lookback period may overestimate the incidence rates [51].

2.2.2 Methods Used to Analyze Varying Lookback Data in Literature

A recent study of Alzheimer's disease (AD) in the US adult cohort showed that influenza vaccination is associated with reduced AD risk for adult people [52]. They used a 6 years lookback period and found a statistically significant relationship between influenza vaccination status and incident dementia. They used propensity score-matched cohorts to find the treatment effect of flu vaccination on AD risk. As they used nearest neighborhood matching without replacement, this can be computationally inefficient, and the treated subjects matched may change the quality of the matches [53]. Another study showed a comparison between the extended lookback period of all available data and a fixed baseline period and concluded that there is a chance of covariate misclassification with a fixed baseline period [20]. In the analysis, three propensity score based Cox-proportional hazard models were used to estimate the outcome. In the first model, they categorized the propensity score into deciles and used it as a covariate in Cox-proportional hazard model. In the second model, an asymmetrical trimming approach was used with 2.5% levels of PS tails. In the third model, nearest-neighborhood matching was used with a 0.025 caliper. However, one problem with these propensity score matching methods arises if there is not substantial overlap between the two groups, in which case bias may be introduced in the model. Another study conducted on US electronic health record database aimed to assess the predictive performance between models with up to 2 years of lookback data [54]. Multiple logistic regression was used for the different combinations of lookback periods and to predict the mortality outcomes, and found no significant clinical differences between the lookback periods. However, they concluded that bias is possible due to missing clinical and demographic information.

2.3 Targeted Maximum Likelihood Estimation (TMLE): A More General Approach to Causal Inference

Targeted maximum likelihood estimation (TMLE) of a parameter of a data-generating distribution is a semi-parametric doubly robust method that improves the chances of correct

model specification by allowing flexible estimation using a (non-parametric) machine learning method. Correct specification of the propensity score model is crucial to obtain an unbiased estimate of marginal treatment effect by matching or inverse probability weighting. The TMLE as a double robust method was proposed by Van der Laan to minimize the impact of model misspecification, which requires estimation of either outcome and treatment mechanisms [22, 55]. Even if one of the models is misspecified, the TMLE effect estimator is asymptotically consistent. Furthermore, TMLE is also locally semi-parametric efficient, meaning that the estimator has (asymptotically) the smallest standard error among estimators that make the same model assumption under the correct specification of both models. In observational data settings, TMLE is a well-established alternative method with desirable statistical properties. A recent simulation study showed how TMLE can be applied in practical settings [56]. Methods based on propensity score are not doubly robust as there is a chance of bias in the estimation of the exposure mechanism; the TMLE can outperform in an observational setting. A comparison was shown between TMLE, G computation, and inverse probability weighting where TMLE had less mean bias compared to the other two methods [56]. The study further demonstrated the best practice of TMLE application using an ensemble machine learning algorithm [56]. Another study by Pang showed a comparison between TMLE and IPTW using a point exposure cohort study of the marginal causal effect in the CPRD data [57]. The study considered the situation of near violations of practical positivity assumption in a high dimensional covariate setting and concluded that TMLE had improved performance in case of well-specification of the outcome model [57]. Another methodological study by Porter and colleagues demonstrated that under the logistic framework, a certain version of TMLE is more robust to violations of positivity assumption and model misspecification than the linear function [58]. Schnitzer and colleagues implemented TMLE methodology to estimate the impact of breastfeeding intervention on gastrointestinal infections [59]. The study showed that TMLE did not produce less bias compared to the G computation substitution estimator, but the TMLE estimators under the logistic framework were more stable in near-positivity violation. The study also showed that the performance of TMLE was well in the time-interval case and better than non-doubly robust methods under model misspecification. Moreover, another article suggested that under model misspecifica-

tion, both TMLE and bias-corrected matching method (BCM) can perform well compared to propensity score methods when the machine learning methods are incorporated [60].

2.3.1 TMLE Using Ensemble Machine Learning Algorithm

In parametric models, biases may arise when the assumed functional form or distribution of the model does not match the true relationship in the data, leading to systematic errors in the estimates [61]. The central limit theorem (CLT) and the law of large numbers are used as classical tools to draw appropriate inferences while working with parametric model approaches [62]. Inference based on these models may produce biased estimates, however, if the model is misspecified. Flexible machine learning tools may be used to alleviate this bias. As TMLE has both outcome and treatment models, machine learning approaches can be used in these models, which incorporate a variety of algorithms [63]. While dealing with observational data, a large number of covariates with potentially complex relationships can be present. So, model misspecification can happen in estimating treatment effects. In this setting, TMLE implementation using machine learning tools can be advantageous. Simulation studies performed by [56] have shown that TMLE using SuperLearner (i.e., ensemble method) helps to reduce bias in average treatment effect (ATE) estimates. Ensemble learning is a technique that uses multiple algorithms (can include traditional regression methods, too) and combine them to advance estimates and predictive performance. There are a variety of ensemble models, i.e., random forest, bagging, boosting, and SuperLearner [63]. Several tutorials aimed at TMLE using SuperLearner have been published in the literature, including R code for implementation as well [56, 64]. In R under library SuperLearner, instead of only simple regression models, a set of prediction algorithms can be used. These algorithms may include parametric regression models, non-linear regression models, shrinkage estimators, and regression trees. The ensemble method SuperLearner selects a weighted combination of different algorithms to optimize cross-validated mean square error [65, 66]. A study evaluated the performance of TMLE using machine learning-based estimators by implementing TMLE using random forest and SuperLearner and concluded that TMLE incorporating SuperLearner produces more feasible estimates with small standard error [67]. Another methodological study has shown that while using the doubly robust

method TMLE, the SuperLearner method worked well with the ensemble including generalized additive model, general linear models, and regression splines [68]. A case-control study on breast cancer was performed to estimate the causal effect of reproductive factors on the risk of breast cancer using an updated version of TMLE (case-control weighted (CCW)) and SuperLearner algorithms [69]. This study identified that TMLE with the SuperLearner, was more efficient for controlling confounding than other statistical approaches. Another study to analyze dietary data has shown the variation of results while using TMLE with logistic regression and ensemble machine learning algorithms [70]. The study concluded that SuperLearner with TMLE produced stronger and more precise estimates with less variation. A simulation study performed by Schnitzer has shown the use of TMLE with SuperLearner in the presence of dependent censoring and intimated that overall TMLE implementation with the SuperLearner library produced the most efficient result [71]. A very recent US study was performed to examine the effect of mobility on new COVID-19 case rates [72]. In the primary analysis, TMLE with SuperLearner was used to adjust confounders, and an unadjusted analysis was also implemented. The unadjusted analysis showed a strong association between mobility indexes and COVID-19 new case rates, while no association was found after confounding adjustment [72]. Finally, they reported that TMLE with SuperLearner is an attractive approach and should be used in an effective way to provide the basis for valid statistical inference [72]. Another study implemented TMLE to estimate the causal association of proximity to gold and copper mines on respiratory disease among the children living in a desert area of Northern Chile. The study concluded that TMLE with ensemble learning algorithms can reduce the parametric misspecification bias and draw valid inferences [73].

2.4 How Much Lookback Length Should be Taken Using Propensity Score Methods?

As noted earlier, in epidemiology, cohort studies usually require a baseline period to measure the effect of covariates as the databases provide varying and truncated historical in-

formation. The accuracy of the covariate estimate and sample size might be affected by the length of this baseline interval. However, a methodological study conducted by Ripamonti et al. (under review in Statistical Methods in Medical Research) has investigated the impact of unmeasured/truncated time-varying covariates on propensity score estimates using different lookback periods. They made a comparison among the types of variables, such as confounders, instrumental variables (iv), and risk factors (rs) in the propensity score model. The simulation study was mainly focused on the effect of unmeasured covariates due to varying lookback windows on propensity score (PS) estimation. The empirical distribution function of propensity score estimates for both the treated and control group were calculated for different lookbacks. The empirical maximum likelihood estimator for minimum, 1^{st} quantile, median, 3^{rd} quantile, and maximum were obtained for various lookback lengths. In the analysis, the lookback lengths of 1, 2, 5, 7, and 9 years were considered, and 10 years was used as a term of comparison. A relevant variation is found in the quality of estimates of the PS distribution as longer lookback time windows guarantee less biased and more accurate estimates of the PS quantiles. From the PS summary, it was expected that the lookback time window of 9 years led to estimates very close to reality. Estimates obtained with lookback windows of 7 years and, in some cases, 5 years generally provided reasonable results both in terms of PS distribution function and outcome model. From all the scenarios of simulation results, estimates obtained with a short lookback time window lead to very poor outcomes. Previous simulation studies suggested that the variables that are unrelated to the exposure but related to the outcome, known as risk factors (rs), should always be included in a PS model [74]. Whereas the variables that are related to exposure but not the outcome, known instrumental variables (iv), should not be inserted in the PS model. A finding from Ripamonti's study suggested that while using a long lookback time window, inserting instrumental variables provided slightly better estimates (under review in Statistical Methods in Medical Research). They explained this situation in two ways. First, with a long lookback, the impact of instrumental variables could be weaker, which may lead to complicated effects such as reducing the estimates of bias or standard errors. Second, they suggested more future simulation studies to consider the correlation between iv and exposure, which is mostly neglected in the literature. Thus an interesting interaction

effect was detected between the type of variable and the length of lookback time windows. By contrast, using shorter lookback windows of one or two years, including instrumental variables (iv), led to inconsistent PS estimates. However, in the literature, it is often believed that the correlation between iv and confounders is weak; however, this relation can be moderate too, in some cases. In that situation, Ripamonti suggested including iv's that may lead to more consistent PS estimates. Thus a careful detection of possible iv's and choice of lookback windows was recommended in this study.

2.5 Knowledge Gap

From the literature review, it is evident that researchers have examined the application of various lookback periods using the propensity score approach. However, the TMLE approach has not been considered with different lookback data. Accurate estimation of treatment effects requires information on an individual's long-lasting medical history, which is often unavailable in registry-based hospital data. Therefore, identifying the optimal lookback becomes crucial to enhance accuracy. As a more general and doubly robust maximum likelihood-based method, TMLE incorporates the targeting step to optimize the bias-variance trade-off in the presence of confounding when estimating the parameter of interest. Surprisingly, no previous study has explored the choice of an optimal lookback length within the TMLE framework. The selection of an optimal lookback length involves a trade-off between the accuracy level of estimating the treatment effect and the number of lookback periods included in the dataset. Furthermore, none of the studies have compared the use of various lookback data with TMLE and PS methods. Such comparisons would provide valuable insights into the strengths and weaknesses of these approaches.

2.6 Objective of the Thesis

- To evaluate the performance of targeted maximum likelihood estimation (TMLE) with varying lookback time windows.

- To compare the effect of metformin vs. sulfonylureas on the risk of major adverse cardiovascular death using TMLE.
- To compare parameter estimation using TMLE and propensity score methods with different lookback time windows in simulation and real life settings.

Chapter 3

Methods

3.1 Potential Outcomes Framework

The Neyman-Rubin causal model is the most widely used potential outcomes framework for causal inferences in statistics and health science [75, 76, 77, 78, 79, 80]. According to this causal model, let us consider a binary treatment indicator A where we specify $A = a$ to a particular treatment assignment. If $a = 1$, we consider the scenario when the individual receives the treatment, and if $a = 0$, we refer to the scenario when an individual does not receive the treatment. We are interested in knowing the effect of A on an outcome variable Y. The potential outcomes framework allows us to expand the joint distribution of (A, Y) to include two random variables $Y(1), Y(0)$. Where $Y(1)$ is the potential outcome if an individual receives the treatment, and $Y(0)$ is the potential outcome if an individual does not receive the treatment as,

$$Y = \begin{cases} Y(1) & \text{if } a = 1, \\ Y(0) & \text{if } a = 0 \end{cases}$$

The potential outcomes are referred to as counterfactual outcomes as they represent what would have been observed in the scenario that did not occur. For example, $Y(0)$ is not observable if $a = 1$ and then $Y(0)$ becomes counterfactual. Similarly, when $a = 0$, $Y(1)$ is the counterfactual. We will discuss the term "counterfactual" briefly in this chapter.

3.1.1 Causal Estimands, Counterfactuals, and Confounding

In evaluating public health policy, the common challenge is to estimate the causal effect of a treatment or intervention. In the realm of statistical inference, the data allows researchers to estimate various types of population causal effects, known as estimands. The term **estimand** is defined as the parameter in a population to be estimated from statistical analysis. The **causal estimands** refer to specific quantities or parameters that represent the causal effects of an exposure or treatment on an outcome of interest in a population [81, 82]. In causal inference, one of the standard problems is causal estimand which is depicted by the array of values in Table 3.1. In Table 3.1, there are N subjects, and each subject is exposed or not exposed to the treatment. The column covariates X represent the variables that take their values for each subject and are independent of the treatment assignment. The column labeled "potential outcomes" represents the value of outcome Y for each subject at a certain time.

Notation: A : Treatment status

$$A = \begin{cases} 1 & \text{treatment, or} \\ 0 & \text{no treatment} \end{cases}$$

Y : Outcome

$$Y = \begin{cases} Y(1), & \text{if under active treatment} \\ Y(0), & \text{if under control group} \end{cases}$$

All of the information should be analyzed by considering X, $Y(1)$, and $Y(0)$. The "subject-level causal effect" represents the comparison of $Y(1)$ and $Y(0)$ for each individual. For any subject i at most one of the $Y_i(1)$ and $Y_i(0)$ can be observed. The last column, "summary causal effect," represents all subjects' mean individual-level causal effects. So, the potential outcomes framework for causal inference can be summarized as the observed outcome is

what actually happened; in contrast, the counterfactual outcome is what would have happened if a different treatment had been assigned [81].

Table 3.1: The Causal Estimand

		Potential outcomes			
Subjects	Covariates X	Treatment $Y(1)$	Control $Y(0)$	Subject-level causal effects	Summary causal effects
1	X_1	$Y_1(1)$	$Y_1(0)$	$Y_1(1)$ vs $Y_1(0)$	
.	.	.	.	.	
.	.	.	.	.	Comparison of $Y_i(1)$ vs $Y_i(0)$ for a common set of units
i	X_i	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$ vs $Y_i(0)$	
.	.	.	.	.	
.	.	.	.	.	
N	X_N	$Y_N(1)$	$Y_N(0)$	$Y_N(1)$ vs $Y_N(0)$	

The **individual treatment effect (ITE)** is a comparison of the potential outcomes for a specific individual i. If Y_{1i} is the potential outcome of an individual i if treated and Y_{0i} is the potential outcome of that individual if untreated, then the individual treatment effect (ITE) will be,

$$ITE_i = Y_i(1) - Y_i(0) \tag{3.1}$$

In this way, we may compare the disease outcome under treatment and no treatment. However, one problem is that we can only observe one outcome (either $Y_i(1)$ or $Y_i(0)$ for any individual i). So, the ITE can never be observed directly, as only one of the potential outcomes may be observed. So this individual effect is not identifiable.

Holland defined this as the "**Fundamental Problem of Causal Inference**" [76]. In this case, an alternative approach is to extend our population where we can evaluate the treatment effects in a particular context but not for a specific individual. This phenomenon brings us to our primary causal parameter of interest, **the average treatment effect (ATE)**. In practice, the causal effect of treatment refers to the specific impact that the treat-

ment has on an outcome of interest, taking into account the cause-and-effect relationship between the two. It aims to quantify how the outcome would change due to the treatment, compared to what would have happened in the absence of the treatment (i.e., the counterfactual scenario) [81, 82].

Under the potential outcomes framework for causal inference, the **average treatment effect (ATE)** is the average of all individual treatment effects in a sample. The ATE can be defined in two ways:

- Average of all differences $Y_i(1) - Y_i(0)$; where $i = 1, 2.....N$
- Difference between average of all $Y_i(1)$ and $Y_i(0)$; where $i = 1, 2...N$

We will describe the estimation and statistical inference for the ATE in **section 3.2**.

Causal inference is drawing the conclusion of the effects of real or conceptual interventions on an outcome. Through the Rubin Causal Model, one can define the causal parameter of interest, and here comes the term confounding. **Confounding** is defined as the exposure and the outcome of interest having a common cause. The bias due to confounding arises when there is a failure to adjust for this common cause. The variable C is called to be a confounder variable when it is causally associated with both exposure X and outcome Y and not on the causal pathway between X and Y (see Figure 3.1).

3.1.2 Assumptions

- **Stable Unit Treatment Value assumption [SUTVA]:** The Stable Unit Treatment Value Assumption (SUTVA) states that the treatment assignments of one unit (e.g., an individual or a group) do not influence the potential outcomes of other units, and there are no hidden treatment variations leading to different treatment effects. [83, 81, 82].

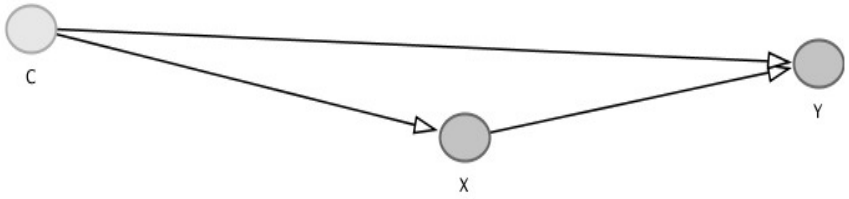


Figure 3.1: DAG representing confounding

- **Consistency:** If an individual or unit receives a specific treatment, then their observed outcome corresponds to the potential outcome under that treatment condition.

- **Positivity:** The Positivity assumption requires that the probability of any individual receiving treatment should be bounded away from zero and one [82].

- **Exchangeability / Strong Ignorability of Treatment Assignment:** The assumption of exchangeability, also known as strong ignorability, posits that the treatment assignment is unrelated to potential outcomes under different treatment conditions, given the observed covariates. In simpler terms, altering the treatment status among the individuals being studied whether from exposed to unexposed or vice versa—will not affect the magnitude or direction of the treatment effect [82, 84].

3.2 Average Treatment Effect (ATE)

The average treatment effect is the mean difference of potential outcomes between subjects assigned to the treatment group and subjects assigned to the control group [82]. When the outcome variable is binary, the average treatment effect (ATE) can be represented as the risk difference (RD). Throughout the entire thesis, we will use the term **ATE** to refer to the **risk difference (RD)**.

Let us assume Y: A binary outcome variable (takes value 0 or 1); A: Treatment variable (binary: 1 if treated, 0 if not treated). So, the risk difference (RD) will be,

$$RD = P(Y(1) = 1) - P(Y(0) = 1)$$

Under the potential outcomes framework, the average treatment effect is defined notationally as follows [82]:
$Y(1)$: potential outcome if an individual receives the treatment ($A = 1$).
$Y(0)$: potential outcome if an individual does not receive the treatment ($A = 0$).
$E[Y(1)]$: expected value of the potential outcome Y(1), which represents the average outcome for individuals if they were all treated ($A = 1$).
$E[Y(0)]$: expected value of the potential outcome Y(0), which represents the average outcome for individuals if they were all untreated ($A = 0$).

Then the average treatment effect (ATE) is defined as:

$$ATE = E[Y(1)] - E[Y(0)] = \sum [Y(1)P(A = 1)] - \sum [Y(0)P(A = 0)] \tag{3.2}$$

In general, we cannot observe both $Y(1)$ and $Y(0)$ since any individual cannot be treated and not treated at the same time. It is probable that the treatment assignment mechanism is dependent on outcomes, leading to potential differences in baseline covariates between the two groups. Such discrepancies in the observed sample means can introduce bias into the estimator of the average treatment effect (ATE). This bias can be seen in the data (overt

bias) or can be hidden [78]. For example, prior to the treatment, the treated subjects might be more financially stable than the non-treated group. A hidden bias occurs when the necessary information is not reported or observed. Overt bias can be adjusted using statistical adjustment [78]. However, the information on baseline covariates can provide useful data for ATE estimation, and we may use this data to find a less biased estimator of ATE. Estimating the ATE requires careful consideration of potential confounding factors.

3.3 Propensity Score Analysis

As described earlier, randomized controlled trials are the gold standard for estimating the causal effect of a treatment. This random treatment assignment maximizes the chance of having treatment groups with the same baseline patient characteristics (known or unknown). However, if there is a discrepancy between the two treatment groups because of different background factors, that may influence both the exposure and the outcome. This caused the phenomenon of "confounding." Confounding is a distortion of the estimated measure due to the presence of a known or unknown variable (confounding variable) associated with both treatment and outcome [85]. Randomization is the ideal approach for estimating causal effects but is not always feasible. Therefore, many statistical techniques have been proposed for non-randomized studies (i.e., observational studies) to reduce the influence of confounding variables. Propensity score methods are one such method. The **propensity score** is the probability that an individual would have been assigned to a particular treatment group as a function of observed baseline covariates [86]. To interpret the propensity score, suppose we have a dichotomous treatment variable $A = (0, 1)$ and a vector of observed available pre-treatment covariates $X = x_1, ...x_n$. Then the propensity score $e(X)$ be the conditional probability of assignment to treatment ($A = 1$), given the covariates X defined by,

$$e(X) = Pr(A = 1|X)$$

where,

$$P(A|X = x_1....x_n) = \prod_{i=1}^{N} e(X_i)^A (1 - e(X_i))^{1-A}; \;\; A \in (0,1)$$

The propensity score is a single scalar value calculated for each individual based on multidimensional covariates denoted as X. It represents the estimated probability of an individual receiving treatment. The objective is to ensure that the distribution of these propensity scores is similar between the two treatment groups. This similarity can be achieved, for instance, by using propensity score matching or inverse probability weighting. As a result, the propensity score acts as a balancing score, as the conditional distribution of covariates, given the propensity scores, becomes comparable between the treated and control groups [86]. Rosenbaum and Rubin showed that if the choice of treatment allocation is strongly ignorable given a set of baseline covariates X, then the propensity score $e(X)$ is a function of these covariates that also makes treatment selection strongly ignorable [86].

As we noted earlier in section 3.2, the assumption of **strong ignorability** is also known as an **unconfoundedness** assumption which states that, the treatment assignment is independent of potential outcomes conditional on observed covariate X. Mathematically, the strongly ignorable (unconfoundedness) assumption can be expressed as follows:

$$Y(1), Y(0) \perp A|X$$

However, the assumption of **exchangeability** states that the potential outcomes $Y(1)$ and $Y(0)$ have the same distribution, given the covariates X. In other words, the potential outcomes are interchangeable between the treated and control groups, conditional on the observed covariates. So, $Y(1), Y(0)$ are exchangeable given X. To show that these two assumptions are equivalent, we will demonstrate that strong ignorability implies exchangeability and vice versa:

Let us assume strongly ignorable such that, $(Y(1), Y(0) \perp A|X)$. Then, consider the following conditional probability:

$$P(Y(1), Y(0)|X, A = 1) = P(Y(1)|X, A = 1) \times P(Y(0)|X, A = 1)$$

Since $Y(1)$ and $Y(0)$ are conditionally independent of A given X, the conditional probabilities are equivalent such that [87]:

$$= P(Y(1)|X, A = 1) \times P(Y(0)|X, A = 1)$$
$$= P(Y(1)|X) \times P(Y(0)|X)$$

The same reasoning applies to the case when $A = 0$

$$P(Y(1), Y(0)|X, A = 0) = P(Y(1)|X) \times P(Y(0)|X)$$

Therefore, the potential outcomes $Y(1)$ and $Y(0)$ have the same distribution given X, which satisfies the exchangeable assumption.

Now the assumption of exchangeability states that, $Y(1), Y(0)$ are exchangeable given X. This implies that the joint distribution of ($Y(1)$, $Y(0)$) does not depend on A, given X and so,

$$P(Y(1), Y(0)|X, A) = P(Y(1), Y(0)|X) \tag{3.3}$$

Then, we can express the conditional probability as:

$$P(Y(1)|X, A)$$
$$= \int P(Y(1), Y(0)|X) dY(0)$$
$$= \int P(Y(1), Y(0)|X, A = 0) dY(0)$$
$$= \int P(Y(1)|X, A = 0, X) P(Y(0)|X, A = 0, X) dY(0)$$
$$= \int P(Y(1)|X) P(Y(0)|X) dY(0)$$
$$= P(Y(1)|X)$$

Similarly, we can show that $P(Y(0)|X, A) = P(Y(0)|X)$. Thus, $Y(1)$ and $Y(0)$ are conditionally independent of A given X, which satisfies the strongly ignorable assumption. Rosenbaum

and Rubin [86] proved that if the potential pair of outcomes $(Y(1), Y(0))$ is not dependent on the treatment assignment given the observed covariates X, they are also independent of treatment allocation given the propensity score $e(X)$:

$$\text{if } (Y(0), Y(1)) \perp A|X \text{ then}$$
$$(Y(0), Y(1)) \perp A|e(X) \text{ and } A \perp X|e(X)$$

Since the propensity score is a balancing score, the mean difference between the treated and control group at a particular value of the propensity score is the average treatment effect at that propensity score (assuming ignorability). Under this assumption, several propensity score methods (defined below) can provide less biased estimates than naive approaches. If the assumptions are not met, bias can still be present in the estimates. Sensitivity analyses and diagnostic checks are often conducted to assess the robustness of results to potential violations of these assumptions [88]. The true propensity score is usually known and designed by study design in randomized control trials. But in observational studies, it is not known; instead, it has to be estimated from study data [13]. These propensity scores can be used in four ways that will be discussed in the next section [18]:

- Matching on propensity score
- Stratification on propensity score
- Covariate adjustment using the propensity score
- Weighting using the propensity score

3.3.1 Matching on Propensity Score

Once the propensity score is estimated, one could match treated to control subjects with similar propensity scores [89]. In the matched sample, differences in outcome between the treatment and control group provide an unbiased estimate of the treatment. There are several types of matching algorithms, such as one-to-one matching, 1:N matching, nearest neighbor matching, N:N matching, caliper matching, radius matching, kernel matching,

Mahalanobis metric matching, etc. **Matching on propensity score** requires a matched pair of treated and control individuals who share a similar value of estimated propensity score. In **nearest neighborhood matching** or **NNM matching**, a treated individual is chosen randomly and then matched to the control individual having a similar or closest propensity score estimate. The first step is to take the absolute minimum difference between estimated propensity scores for the control and treated individuals. Then the treated and control units are randomly ordered such that the first treated unit is matched with a control unit having the closest propensity score estimates.

$$C(P_i) = \min_j |P_i - P_j|$$

Here, $C(P_i)$ is defined as the group of control participants j matched to the treated subjects i (matched on the estimated propensity score)

P_i is the estimated propensity score for the treated individuals i

P_j is the estimated propensity score for the individuals in control group j

This can be done "with replacement" or "without replacement". In the "with replacement" approach, an individual in the control group can be used more than once for an individual in the treated group. For the "without replacement", the control group's comparison unit can be considered a match with the treated units only once. However, matching with replacement works as a bias-variance trade-off as if the replacement is allowed, the average matching quality will increase, and hence the bias will decrease. This NNM matching is also known as **greedy matching** as in each stage, the control unit is selected close to the currently considered treated unit, even if the untreated unit would work better as a match for a subsequently treated unit. However, one problem may arise with NNM matching when the close neighbor is far away. This can be avoided by imposing a pre-determined range of values usually defined within one-quarter of the standard error ($0.25s$) of the estimated propensity score. This is called **caliper matching**, where the control unit is chosen as a

match for a treated unit within the caliper (pre-determined range). However, one problem with this matching method is not knowing the choice of the possible range in advance [90].

Another matching method as an alternative to caliper matching is **radius matching** recommended by [91]. In this approach, every treated participant is matched with an analogous control participant that lies within a pre-determined range of the treatment unit's propensity score. Thus a comparison is made among all the caliper members. The method might have the drawback of oversampling, but the risk of bad matches can be avoided. Another non-parametric matching method is known as **kernel matching**, where individuals in the treated group are matched with individuals in the control group based on their estimated propensity scores [92]. The matching process assigns weights to each control group member according to a kernel function that captures the similarity of their propensity scores to the treated individuals [92]. These weights are inversely equivalent to the distance between treated and control group units. In **N:N matching** method, control and treated units are randomly assigned, but the first n treatment units are matched with n control units with the adjacent propensity score. Some other propensity score matching methods are $1:1$, $1:N$ or $N:1$ matches, Mahalanobis metric matching, etc. [89].

3.3.2 Stratification or Subclassification on the Propensity Score

In stratification, the estimated propensity score divides the population into homogenous subclasses with similar propensity scores [14, 93]. This method is commonly applied in observational studies to control systematic discrepancies between treated and control units. At first, the participants are grouped according to their estimated propensity score. A subset/subclass of participants is created based on the formerly defined threshold of the estimated propensity score. Thus the strata are defined, and a comparison is made between the treated and control units in each stratum. The treatment effect is estimated within each stratum, and finally, the treatment effects for all strata are combined to estimate the overall treatment effect. Several examples are found in literature where both regression adjustment and using propensity score strata as a covariate can be used in the model [94, 95, 96]. Following Cochran, these subclasses are mostly based on percentiles, quintiles, etc

[97]. In a seminal article, [97], Cochran concluded that 90% of the covariate bias could be reduced by creating quintiles of a continuous confounding variable. However, by classifying these groups on quintiles, Rosenbaum and Rubin [86] demonstrated that approximately 90% of bias due to measured confounders can be reduced while estimating linear treatment effect. In an interview with observational studies [98], Rubin stated that a sufficient number of strata (i.e., five sub-classes per thousand subjects) could ensure a more negligible bias than a variance.

3.3.3 Inverse Probability of Treatment Weighting (IPTW) using Propensity Score

Rosenbaum initially put forth the method of model-based direct standardization, while the inverse probability weighting (IPW) technique was introduced by Donald B. Rubin in 1980 and later elaborated upon by James M. Robins in his influential paper titled "Causal Inference from Complex Longitudinal Data" published in 1997 [18, 99, 100]. Under the assumptions of consistency, exchangeability, positivity, and correct specification of the model, the inverse probability treatment weighting can reduce the confounding by creating a pseudo-population [82]. This pseudo population is generated by assigning a weight to each individual that is proportional to the treatment option they actually received [82]. Under exchangeability, the exposed and unexposed groups are, on average, similar, conditional on covariates. This assumption implies that there are no unmeasured confounders and residual confounding and makes a fair comparison between the two groups. However, in observational research, this is unrealistic as we can never observe the unmeasured confounding and can only adjust the measured confounders and hence only "conditional exchangeability" can be assumed [101, 102]. The inverse probability weighting method utilizes the estimated propensity score values to construct sets in which the treated and control groups possess, on average, similar characteristics. For the treated groups, weights for each individual are calculated as 1/(propensity score) and for the control group 1/(1 – (propensity score)). After that, these weights are assigned to balance individual baseline characteristics for the treatment and control groups. Finally, in the pseudo-population, the outcome variable is re-

gressed on the exposure by fitting a crude weighted regression model [101]. The estimated parameters from the regression model are used to estimate the causal effect of the exposure in the study population. Again, denoting e_i the propensity score for ith subject.

$$e_i = P(A_i = 1|X_i)$$

The most widely used model for estimating propensity score is logistic regression [53, 16]. Since the treatment indicator A_i is binary and X_i denotes the vector of measured baseline covariates. The logistic regression model is parametrized by, $\alpha = (\alpha_0, \alpha_1, ...\alpha_p)^T$ such that,

$$log\left(\frac{ei}{1-ei}\right) = {X_i}^T \alpha$$

For each individual indexed by subscript i, the probability of being treated or not treated given the baseline covariates is estimated from the fitted propensity score model as [23]:

$$\hat{e_i} = \frac{exp(X_i^T \hat{\alpha})}{1+exp(X_i^T \hat{\alpha})}$$

Then for the ith treated individual, the weight is,

$$w_i = \frac{A_i}{\hat{e_i}}$$

And for ith control individual, the assigned weight,

$$w_i = \frac{1-A_i}{1-\hat{e_i}}$$

Finally, the weight for ith individual,

$$w_i = \frac{A_i}{\hat{e_i}} + \frac{1-A_i}{1-\hat{e_i}}$$

Thus an individual's weight is equal to the inverse of the probability of receiving the treatment that the individual actually received. So, individuals in the treated group with a lower probability of treatment are likely to receive larger weights. We will illustrate this phenomenon with the following example.

Suppose there are two individuals, A and B. A researcher estimates the probability of being treated for A and B as 0.5 and 0.8, respectively. The weights are calculated as 1/0.5 = 2

and 1/0.8 = 1.25. So, person A would be given more weight than B. Thus B is given smaller weights as there can be a reason why B is more likely to be treated. Similarly, the untreated individual having a higher probability of being treated (lower probability of being untreated) receives higher weight. Thus the data become more comparable and similar. In some cases, however, this method can produce biased results. If the model specification is wrong, this covariate balance might get worse. In this case, doubly robust estimation is popular for solving issues, as discussed in a later section.

3.3.4 Covariate Adjustment on Propensity Score

This method includes using the estimated propensity scores as an additional variable in the regression model. Under this approach, the outcome variable of interest is regressed on the exposure/treatment variable and the estimated propensity score [16]. The regression model is chosen depending on the outcome of interest. For instance, a simple linear regression model can be used for the continuous outcomes, or a logistic regression model can be chosen for a binary outcome. The treatment effect is defined from the coefficient of a fitted regression model. In contrast, using propensity score as a covariate in a multivariable model, researchers cannot take the full convenience of the propensity score features. In an article by Austin, it was explained that this method assumes the same distributional assumption of the baseline covariates between the treated and untreated group [103]. So, covariate adjustment can be easily affected by the assumptions of distributions and correct propensity score specification. Moreover, differences in the variance of the propensity score function between the treated and untreated groups may also lead to bias [103]. However, researchers need to be careful with the analysis while using this method as it is not considered good practice [104]. Besides, it does not permit the assessment of covariate balance like the other three methods (i.e., matching, stratification, IPTW).

3.4 Average Treatment Effect Estimation using Inverse Probability Weighting

The individual weights described in section 3.3.3 are used to create a weighted sample where no confounding is present in the observed baseline covariates. To estimate the average treatment effect (ATE), one can easily use this weighted sample. Let us assume a binary treatment indicator (where $A = 1$ and $A = 0$ for the treatment and control group, respectively). The propensity score for each individual is defined as $e = P(A = 1|X)$; the probability of an individual receiving treatment given the observed baseline covariates [86]. As described earlier, the individual treatment weight is defined by $w = \frac{A}{e(X)} + \frac{1-A}{1-e(X)}$. The inverse probability weighted mean is estimated using the Horvitz-Thompson estimator [105] and by definition of ATE [82],

$$\tau^{ATE} = E\left[\frac{AY}{e(X)} - \frac{(1-A)Y}{1-e(X)}\right]$$

For ith individual the inverse probability weight is,

$$\begin{cases} w_1(X_i) = \frac{1}{e(X_i)}, & \text{for } A_i = 1 \\ w_0(X_i) = \frac{1}{1-e(X_i)}, & \text{for } A_i = 0 \end{cases}$$

Then an unbiased nonparametric estimator of ATE can be expressed by taking the differences between the mean of weighted outcomes between groups as,

$$\tau^{ATE}_{ipw,1} = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{(1-e(X_i))} = \frac{1}{n}\sum_{i=1}^{n} Y_i A_i w_1(X_i) - Y_i(1-A_i)w_0(X_i) \qquad (3.4)$$

It is recommended by [100] to use the "stabilized weights" by diving each individual's weight by the sum of all weights in that group $w_i/\sum_{i:Z=z} w_i$ for $A = 0, 1$ and for instance, the Hajek estimator:

$$\tau_{ipw,H} = \frac{\sum_{i=1}^{n} Y_i A_i w_1(X_i)}{\sum_{i=1}^{n} A_i w_1(X_i)} - \frac{\sum_{i=1}^{n} Y_i(1-A_i)w_0(X_i)}{\sum_{i=1}^{n}(1-A_i)w_0(X_i)}$$

Both stabilized and non-stabilized inverse probability weights lead to the same estimate. But using stabilized weights can result in narrower 95% confidence intervals than non-stabilized weights [82] particularly in the case when the model is not saturated (a model "saturated" means that the model is complex and fully accounts for all available information and covariates). However, another estimator of ATE was reviewed by Lunceford, [23]

$$\tau_{ipw,2}^{ATE} = \left(\sum_{i=1}^{n} \frac{A_i}{ei}\right)^{-1} \sum_{i=1}^{n} \frac{A_i Y_i}{e_i} - \left(\sum_{i=1}^{n} \frac{1-A_i}{1-ei}\right)^{-1} \sum_{i=1}^{n} \frac{(1-A_i)Y_i}{1-e_i} \tag{3.5}$$

Here, equation 3.5 is known as **ratio estimator**. This estimator is often used with IPW to handle missing outcome data and helps in estimating the population-level parameter while adjusting for the non-response or missingness [23]. In the given equations, n represents the total number of individuals, and ipw stands for inverse probability weighting. Specifically, the approaches denoted as $ipw,1$ and $ipw,2$ are two commonly used methods for weighting in the context of causal inference [23].

Before estimating the causal effect, it is recommended to check the standardized mean difference (SMD) between two groups [106]. This can be evaluated before and after weighting allowing for checking balance across the measured variables. The standard error of our parameter of interest ATE can be obtained using various methods. The most ordinary approach is to use a **naive variance estimator** that is based on the assumption of the statistical model (usually logistic regression model) used in estimation [107]. Nevertheless, the model-based standard errors can be incorrect because of the violation of **independent and identically distributed or iid** assumption [18]. This complication arises as the weights are estimated from the data; sampling variability is possible. Since the weights are not fixed, a biased standard error can occur. So, it is recommended to use **robust sandwich estimator standard errors** [108, 109, 110]. Also, bootstrap standard errors can be used

where the approximation is based on a sampling distribution through repeated sampling. In our analysis, we will use **robust sandwich estimator standard errors** under the PS-IPW approach to minimize the chance of biased standard errors.

3.5 Overview of Targeted Maximum Likelihood Estimation (TMLE)

Targeted maximum likelihood estimation (TMLE) is a doubly robust technique that minimizes bias and targets the parameter of interest by leveraging the influence function [22]. An influence function measures an estimator's sensitivity to data perturbations, and TMLE employs this information to update treatment and outcome models iteratively [111, 22]. In general, TMLE benefits from the **efficient influence function** (a type of influence function), which optimizes efficiency and asymptotic properties by achieving the Cramér-Rao lower bound and providing the minimum variance property in the class of semi-parametric estimators of the interesting parameters [22, 112, 113]. Moreover, the inference is drawn for an estimator corresponding to the efficient influence function would be optimally efficient in this class of estimators [114]. Typically, the TMLE estimators are two-stage estimators.

In the first stage, the researcher needs to define the parameter of interest as a function of the data-generating distribution. Usually, practitioners assume the parameter to be a regression coefficient in this stage. However, in many applications, the target quantity is not easily expressed as a model parameter. In TMLE, the estimation procedure involves minimizing a targeted loss function [115]. A targeted loss function combines the log-likelihood of the outcome model and the balancing constraint, ensuring that the treatment assignment mechanism is balanced between treated and control groups [22, 115]. Integrating TMLE with SuperLearner, utilizes this loss-based estimation concept and enhances the efficiency of the estimates, and assures adaptivity to the data at hand during the estimation procedure [22, 116]. Thus the initial estimator from this stage is somehow already informed about the target parameter of interest, focusing on the set of all possible probability distributions of

the data. So, the true probability distribution falls within the assumed statistical model to represent true knowledge.

The second stage of TMLE is referred to as the **targeting step**, where the primary goal is to enhance the initial estimates obtained in the first stage. This phase emphasizes refining the estimation process specifically for the target parameter of interest, employing a targeted maximum likelihood approach. Here, the initial estimator serves as an "offset," and the fluctuation function is applied to this offset. The term **offset** refers to the use of the initial estimated propensity scores as a fixed reference value to update the outcome model predictions. On the other hand, the fluctuation function refers to the difference between the predictions of the outcome model using the updated propensity scores and the predictions from initial propensity scores. Thus a set of potential fluctuations is generated representing the various parametric variations of the offset and identifies a parametric model consisting of fluctuated versions of the offset. This parametric model is a minimally favorable parametric model such that its maximum likelihood estimator listens approximately to the data with respect to fitting the target parameter as a semiparametric model efficient estimator. Thus the second stage is targeted toward making an **optimal bias-variance trade-off** for the parameter of interest rather than the overall distribution [22].

The procedure of the above two stages is called **double robust**. This notable feature **doubly robust** of TMLE means that if either the model for expected outcome ($E(Y|A,X)$) or the model to estimate the probability of treatment $P(A = 1|X)$ is correctly specified the final TMLE estimate will be consistent. Asymptotically if the model and other nuisance parameters are correctly specified, the estimate would be maximally efficient [22]. Furthermore, this property provides **robustness against misspecification** in either model and enhances the validity of causal inference in observational studies. The above-mentioned "targeting" step allows the final ATE estimate to remain consistent if the outcome model is not correctly estimated. The unbiasedness will still be preserved if the treatment model is correctly estimated. An appealing property of TMLE estimators is "robustness," as these estimators are robust to outliers [22]. Another attractive feature of TMLE is **"dealing with**

missingness" described by Van der Laan [117]. The authors stated that TMLE could inherently adjust for dropout (missingness) and can be used to evaluate the treatment effect in non-compliance (when the patient refuses to take a medication). In addition, TMLE can accommodate missing data and censoring by using appropriate imputation techniques, inverse probability weighting, or censoring adjustment methods.

Observational studies often include a large number of covariates. TMLE is well-suited for situations with high-dimensional confounding in the presence of many potential confounders. Traditional propensity score methods may face challenges in such scenarios due to the "curse of dimensionality" and the need for dimensionality reduction techniques. TMLE can handle high-dimensional data by leveraging modern machine learning algorithms, such as lasso or elastic net regularization, to select relevant confounders and estimate the treatment effect more efficiently [118]. Many covariates may exist while dealing with observational data, and a complex relationship might happen between these variables. In that case, model misspecification, which can be a specific concern, can be possible. These models can be misspecified such that the effect estimate can be biased. In addition, TMLE enables the integration of advanced machine learning algorithms into the estimation process. TMLE enables sophisticated machine learning algorithms like random forests, support vector machines, or neural networks for estimating propensity scores or outcome models. This flexibility empowers researchers to leverage advanced techniques for propensity score estimation and outcome modeling within the TMLE framework. In practice, the researcher does not know which machine learning algorithm should be used. Ensemble methods such as **SuperLearner** [119] allows users to specify several machine learning methods. By adopting a comprehensive library approach, the combination of TMLE and SuperLearner facilitates the incorporation of versatile machine learning algorithms [25, 22, 66].

3.5.1 Illustration of TMLE (A Step-by-Step Algorithm)

We illustrate TMLE for estimating the marginal causal effect of a binary point treatment A on observed binary outcome Y. In the description of the TMLE approach, we denote W as a vector containing all important covariates to align with conventional notation in literature

(this remains consistent with X in our previous discussion). Given a full (counterfactual) dataset containing observations where $Y(1)$ corresponds to the outcome observed when subject i is assigned to treatment group ($A = 1$) and $Y(0)$ corresponds to the outcome observed when subject i is assigned to control group ($A = 0$). Thus TMLE can be used to estimate the proportion of subjects experiencing a certain event if everyone were treated in a target population and the proportion with such event if everyone were untreated. We refer to these two quantities as our parameter of interest and the average treatment effect (ATE) can be summarized by using these two corresponding parameter estimates. So, the targeted parameter of interest ATE can be expressed as,

$$ATE = E[Y(1)] - E[Y(0)]$$

TMLE is implemented in the following four steps.

Step 1: In the first step, an initial model for the outcome Y is fit to all observations, using confounders and treatment as a predictor. For a binary outcome, the standard logistic regression model is one possible approach such that,

$$\begin{aligned} &logit([P(Y = 1|A, W)]) \\ &\quad = \alpha_0 + \alpha_1 + {\alpha_2}^T W \end{aligned} \tag{3.6}$$

Using this model, the initial probability for the original dataset can be estimated by:

$$\hat{Q}^0(A, W) = expit(\hat{\alpha_0} + \hat{\alpha_1} + \hat{\alpha}_2^T W) \tag{3.7}$$

The predicted probability for both counterfactual outcomes for treatment group ($A = 1$) and control group ($A = 0$) for everyone are respectively:

$$\hat{Q}^0(1,W) = expit(\hat{\alpha_0} + \hat{\alpha_1} + \hat{\alpha}_2^T W)$$
$$\hat{Q}^0(0,W) = expit(\hat{\alpha_0} + \hat{\alpha}_2^T W)$$

Step 2: In this step, the probability of receiving treatment for all individuals is estimated using the covariates as confounders. Generally, it is known as the propensity score model. A logistic regression model can be used such that:

$$logit[P(A=1|W)] = logit(p_{A|W}) = \beta_0 + {\beta_1}^T W \tag{3.8}$$

The probability of A given W can be estimated by:

$$\hat{g}(1|W) = expit(\hat{\beta}_0 + \hat{\beta}_1^T W)$$

Step 3: In this step, the information about the treatment mechanism in step 2 is used to optimize the bias-variance trade-off for ATE estimate to obtain a valid inference. This is usually done by solving an equation to determine how much to update or fluctuate our initial estimates. To find a better prediction model targeted at minimizing mean square error for the estimation of $u_0 = E[Y(A=0)]$ and $u_1 = E[Y(A=1)]$ by the so-called efficient fluctuation of u_1 and u_0. To solve the equation, a parametric model is defined as:

$$\begin{aligned} logit(E[Y|A,W]) &= logit(\hat{E}[Y|A,W]) + \epsilon_0 H_0(A,W) + \epsilon_1 H_1(A,W) \\ &= \hat{\alpha_0} + \hat{\alpha_1} + \hat{\alpha}_2^T W + \epsilon_0 H_0(A,W) + \epsilon_1 H_1(A,W) \end{aligned} \tag{3.9}$$

where ϵ_0 and ϵ_1 are so-called fluctuation parameters. Hence, $H_0(A,W) = \frac{I(A=0)}{\hat{g}(0|W)}$ and $H_1(A,W) = \frac{I(A=1)}{\hat{g}(1|W)}$ are called clever covariates. The inverse probability of receiving treatment $\hat{g}(1|W)$ and the negative inverse probability of not receiving treatment $\hat{g}(0|W)$ is estimated from the second step. The indicator functions $I(.)$ take value one if one of its Boolean

arguments is true and value zero otherwise.

Thus, fitting a logistic regression solves an efficient influence function ($u0, u1$) estimating equation that satisfies many useful statistical properties of TMLE, such as:

- A long as either outcome or treatment is estimated correctly (consistently), the final estimate is consistent.
- If both are estimated correctly, the final estimate achieves its smallest possible variance as the sample size approaches infinity (efficiency).

The above parametric model is fitted with maximum likelihood estimation to obtain estimates for the fluctuation parameters ($\hat{\epsilon}_0, \hat{\epsilon}_1$). This fitting can be done with standard software by setting expected outcomes under observed treatment,

$$logit(\hat{Q}^0(A,W)) = \hat{\alpha_0} + \hat{\alpha_1} + \hat{\alpha}_2^T W$$

as an offset in an intercept-free logistic regression with covariates H_0 and H_1. More specifically,

$$logit(\hat{Q}^1(A,W)) = logit(\hat{Q}^0(A,W)) + \epsilon_0 H_0(A,W) + \epsilon_1 H_1(A,W) \tag{3.10}$$

Then by taking the model coefficients, the estimates of the fluctuation parameter ($\hat{\epsilon_0}, \hat{\epsilon_1}$) are obtained.

Thus the magnitude of the fluctuation parameter reflects the strength of the association of a function of the propensity score with the signal in the residuals. The value of the fluctuation parameter will be close to 0 if there is a little signal in the residuals.

By replacing ($\hat{\epsilon_0}, \hat{\epsilon_1}$), the estimated probability of Y given A and W and the probability of the counterfactual outcomes for each individual can be updated as :

$$\hat{Q}^1(A,W) = expit(logit(\hat{Q}^0(A,W)) + \hat{\epsilon}_0 H_0(A,W) + \hat{\epsilon}_1 H_1(A,W))$$

So, we can compute the updated outcome under treatment,

$$\hat{Q}^1(1,W) = expit(logit(\hat{Q}^0(1,W)) + \hat{\epsilon}_1 H_1(1,W)$$

and the updated outcome under no treatment,

$$\hat{Q}^1(0,W) = expit(logit(\hat{Q}^0(0,W)) + \hat{\epsilon}_0 H_1(0,W).$$

This updating is done by placing $A = 0$ and $A = 1$ for each individual in the probability functions $\hat{Q}^0(0,W), \hat{Q}^0(1,W)$ similarly as in the clever covariates $H_1(1,W)$and $H_0(0,W)$. As a general rule, TMLE is an iterative procedure, where $\hat{Q}^k(A,W)$ is updated with $\hat{Q}^{k+1}(A,W)$ and above updating is done until convergence (until $\hat{\epsilon}$ is adequately small). The subscript k is used to denote updating the kth step. Generally, the targeting step only estimates the "treatment" mechanism $\hat{Q}^1(A,W)$. In the next step, the average treatment effect (ATE) is calculated by using the above-mentioned updated outcome pairs. In the ATE estimation, it can be demonstrated that the convergence of TMLE can be achieved after a single iteration [25].

In the final step, the ATE is calculated by taking the average difference between the updated expected outcomes across individuals using the G computation formula [120]. G computation formula is an analytic tool for estimating standardized ATE concerning a "standardized" population distinguished by marginal covariate distribution [121, 122, 56]. This estimation technique uses the G computation formula to estimate the average treatment effect (ATE) after adjusting the confounders [123].

The ATE using targeted maximum likelihood estimation is calculated by taking the average differences between the updated expected outcome (via G computation formula [123]):

$$ATE_{TMLE} = \hat{u_1} - \hat{u_0} = E_W\left[\hat{Q}^1(1,W)\right] - E_W\left[\hat{Q}^1(0,W)\right] = \frac{1}{n}\sum_{i=1}^{n}\hat{Q}^1(1,W_i) - \frac{1}{n}\sum_{i=1}^{n}\hat{Q}^1(0,W_i)$$

After performing targeted maximum likelihood estimation, the marginal odds ratio can be obtained. The **marginal odds ratio** represents the change in the odds of the outcome variable associated with a one-unit change in the treatment variable while considering the average effect across the entire study population. So, the marginal odds ratio (OR) can be estimated by the following equation:

$$\widehat{MOR}_{TMLE} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{Q}^1(1,W_i)\right)\left(1-\frac{1}{n}\sum_{i=1}^{n}\hat{Q}^1(0,W_i)\right)}{\left(1-\frac{1}{n}\sum_{i=1}^{n}\hat{Q}^1(1,W_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n}\hat{Q}^1(0,W_i)\right)}$$

3.5.2 Statistical Inference Under TMLE Framework:

The "targeting" step in the TMLE procedure primarily relies on the notion of the efficient influence curve (EIC) [58, 25]. The **efficient influence curve** tells us how much the estimate of average treatment effect would change if the data were slightly modified, preserving the underlying distribution of the data [25]. Following semi-parametric and empirical process theory, an estimator is asymptotically linear if :

$$\widehat{ATE} - ATE = \frac{1}{n}\sum_{i=1}^{n} IC(O_i) - o_p\left(\frac{1}{\sqrt{(n)}}\right)$$

Using the weak law of large numbers, the term $o_p\left(\frac{1}{\sqrt{(n)}}\right)$ will always converge to 0 in probability as the sample size (n) goes to infinity [22, 64]. To be efficient, the researcher must ensure that the influence curve has finite variance where the empirical mean of the influence curve (IC) is 0. Many influence functions may exist in a given estimation problem and for a targeted parameter, and the most "efficient" IC achieves the lower bound on asymptotic variance.

Explicitly, the efficient influence curve for the ATE parameter can be defined as:

$$EIC_{ATE} = \left(\frac{A}{P(A=1|W)} - \frac{1-A}{P(A=0|W)}\right)[Y - E(Y|A,W)] + E(Y|A=1,W) - E(Y|A=0,W) - ATE \tag{3.11}$$

This can be estimated as,

$$\widehat{EIC}_{ATE} = H(A,W)\left[Y - \hat{Q}^1(A,W)\right] + \hat{Q}^1(1,W) - \hat{Q}^1(0,W) - \widehat{ATE}_{TMLE}$$

for every individual. A more sophisticated explanation of the above formula can be found in Chapter 5, Appendix A of [22]. Finally the standard error of $\widehat{ATE}_{TMLE}$ can be defined as follows:

$$\hat{\sigma}_{ATE,TMLE} = \sqrt{\frac{\widehat{Var}\left(\widehat{EIC}_{ATE}\right)}{n}}$$

In which the term $\widehat{Var}\left(\widehat{EIC}_{ATE}\right)$ is defined as the sample variance of the estimated Influence function. The corresponding Wald-type 95% confidence interval for ATE can be constructed by :

$$\widehat{ATE} \pm 1.96\sqrt{\frac{\widehat{Var}\left(\widehat{EIC}_{ATE}\right)}{n}}$$

And the causal null hypothesis $H_0 : ATE = 0$ can be tested with the statistic [22]

$$T = \frac{\widehat{ATE}}{\sqrt{\frac{\widehat{Var}\left(\widehat{EIC}_{ATE}\right)}{n}}}$$

Defining the two quantities D_1 and D_0, the efficient influence curve (EIC) for $E[Y(1)]$ and $E[Y(0)]$ can be expressed accordingly :

$$D_1 = \left(\frac{A}{P(A=1|W)}\right)[Y - E(Y|A=1,W)] + E(Y|A=1,W) - E[Y(1)]$$

$$D_0 = \left(\frac{1-A}{P(A=0|W)}\right)[Y - E(Y|A=0,W)] + E(Y|A=0,W) - E[Y(0)]$$

Using the functional delta approach [124], the efficient influence curve (EIC) for marginal odds ratio is [64]:

$$EIC_{MOR} = \frac{1 - E[Y(0)]}{(1 - E[Y(1)])^2 \times E[Y(0)]} \times D_1 - \frac{E[Y(1)]}{(1 - E[Y(1)]) \times (E[Y(0)])^2} \times D_0$$

3.5.3 Targeted Maximum Likelihood Estimation (TMLE) with SuperLearner

TMLE using SuperLearner gives additional opportunities to minimize the prediction error. SuperLearner uses the output of different machine learning methods that are more flexible than parametric methods. In section 3.5.2, we mentioned that TMLE uses the influence function that requires the models for expected observed outcome $E[Y|A,W]$ and propensity scores $E[A|W]$ to converge to the true value [58, 25]. When the true propensity score model is logistic, the maximum likelihood estimates (MLE) converge to the true value. But in practice, the true model is unknown. In that case, SuperLearner takes advantage of using its candidate models to have the best chance of approaching the true value. Incorporating SuperLearner in TMLE allows combining a set of regressions to construct a library of algorithms containing all weighted averages of these regressions [22]. One of these weighted averages might work better than one of these regression models alone. Thus instead of using the single parametric model for both treatment and outcome mechanisms, the SuperLearner (SL) is implemented to use the information from the data itself flexibly and adaptively. Moreover, SuperLearner has been referred to as the best choice among the family of weighted combinations of targeted estimators [22, 66, 125]. The SuperLearner algorithm requires combining multiple estimators into an improved estimator and gives back a function that can be used to predict new datasets [22].

In order to explain how SuperLearner works with TMLE, we need to define some notation and describe the process step by step as follows:

- Y: The binary outcome variable takes 0 or 1
- A: The binary treatment variable takes 0 or 1
- W: A vector of covariates or confounding variables.
- $g(A)$: The propensity score model that estimates the probability of treatment assignment given covariates W

- $Q(W)$: The outcome model that estimates the probability of the outcome given covariates W
- $g_0(A)$ The true (but unknown) treatment probability model
- $Q_0(W)$ The true (but unknown) outcome model.

In step 1, the SuperLearner algorithm selects a set of candidate algorithms for each model. Some commonly used algorithms for machine learning methods in the context of SuperLearner are logistic regression, decision trees, random forests, neural networks, etc [66]. For each candidate algorithm, cross-validation is used to fit the model to the data and assess its performance [126]. Cross-validation is a statistical technique that is commonly used in machine learning to evaluate the performance and generalization ability of a predictive model on an independent dataset [126]. It divides the data into two parts: a training set and a testing (validation) set. The model is trained on the training set, and its performance is assessed on the testing set [126, 127].

In step 2, the candidate models are combined using a weighted average or other ensemble learning techniques [66, 22]. These weights are computed based on the performance of candidate models during the cross-validation. The resulting SuperLearner models $Q_{SL}(W)$ and $g_{SL}(A)$ are the combined outcome model and treatment model. The treatment probability model $g_{SL}(A)$ predicts the probability of receiving treatment given the covariates W. In the 3rd step, using the estimated treatment probabilities from these SuperLearner models, the inverse probability weights are calculated as:

$$\frac{A}{g_{SL}(A)} + \frac{1-A}{1-g_{SL}(A)} \tag{3.12}$$

Finally, these weights are incorporated in the TMLE approach (see section 3.5.1), and the treatment probabilities and outcome models are updated. After the TMLE procedure, the final causal effect is estimated similarly as described in section 3.5.1.

In this thesis, for our analysis purpose, we will use the following SuperLearner (SL) functions [128] that can be called from library **SuperLearner** under statistical software **R**:

- SL.glm : conventional main terms logistic regression
- SL.step: stepwise regression
- Sl.gam: generalized additive models
- SL.glm.interaction: logistic regression variant that includes second-order polynomials and pairwise interactions of the main terms
- SL.glmnet: lasso (least absolute shrinkage and selection operator)
- SL.rpart: an implementation of classification and regression trees.[129]
- SL.randomForest: implements Breiman's randomForest algorithm [130]

In R, under the package **"tmle"** and **"SuperLearner"**, we will use the above selected SL library instead of the default SL library.

Chapter 4

Simulation Study

4.1 Simulation Settings

Following the data generation setting used by Ripamonti et al. (under review at Statistical Methods in Medical Research), we performed a Monte Carlo simulation study to examine how the lookback period affects the estimation of the average treatment effect using TMLE. We generated datasets focusing on a realistic scenario. Setting an ideal timeline with an initial point of follow-up (t_0), we theorized to have all available information in an entire lookback period with length of 10 years ($t_0 - t_{-10}$). Within this period, two time-varying confounders representing the role of two acute events indicated by c_1 and c_2 were generated from Bernoulli(.50) and Bernoulli(.25) respectively. Another confounder c_3 was generated from Bernoulli (.25) to indicate the emergence of a chronic condition. Further, a time-varying covariate specifying acute events c_4 was generated from Bernoulli (0.5). In addition, an instrumental variable (iv) and an independent risk factor (rs) were generated from Bernoulli (0.50). All these variables were assumed to appear up to three times in ten years. We assumed a moderate association (an odds ratio of 2) between covariate and treatment (or outcome). In our data generation setting we kept all confounders (c_1, c_2, c_3, c_4) along with instrumental variable (iv) and risk factor (rs) with a moderate association. An **instrumental variable (iv)** is a factor that is assumed to be related to exposure but not directly related to the outcome or not indirectly connected via pathways through unmeasured variables. Typically, an instrumental variable depends on three assumptions [131]:

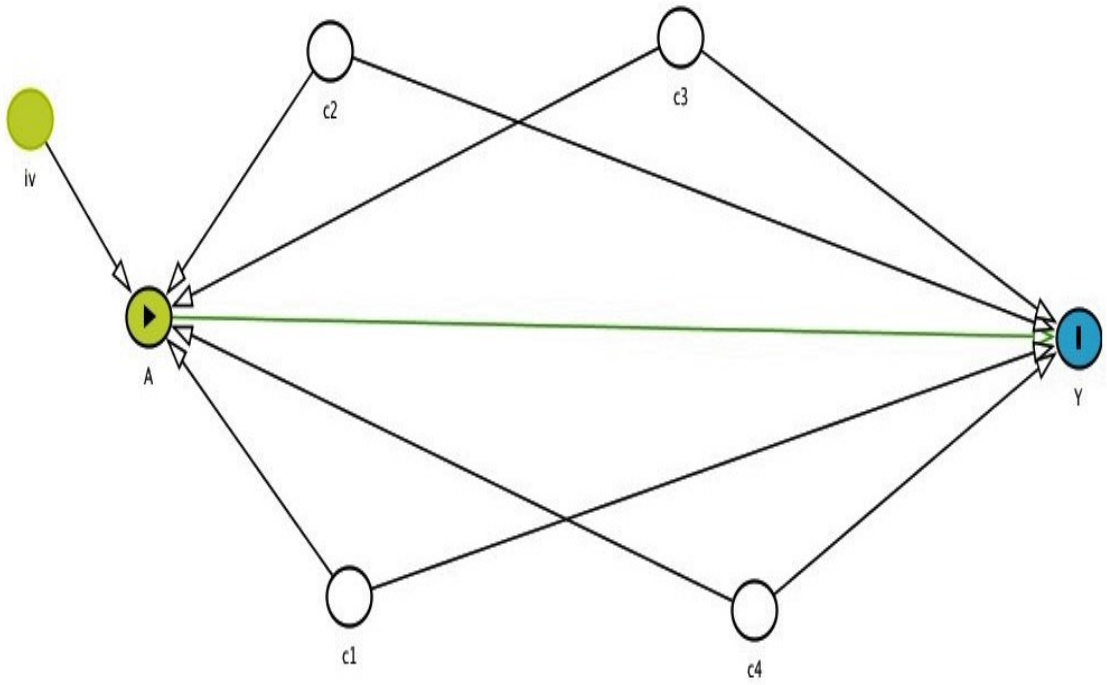


Figure 4.1: Causal diagram representing confounding variables

a) An instrumental variable (iv) is an additional variable that affects exposure or is associated with exposure by sharing a common cause. b) Instrumental variable (iv) is not independent of exposure, but it affects outcome only through its association with exposure [132] (page 247). c) An instrumental variable is independent of all baseline variables, so it is unrelated to patient characteristics. Under observational settings, using large-scale databases, it is often difficult to find proper instrumental variables, but it depends mostly on knowledge, experience, and the researcher's intuition [133]. However, instrumental variables can be used in several settings. For example, an instrumental variable can have a direct association with medical interventions that may influence the outcome, which can be used to estimate the causal relationship between an exposure and outcome. The **risk factor** (rs) is a variable associated with an increased risk of developing the disease or outcome. A risk factor can be behavioral, physiological, demographic, environmental, or genetic [134].

In our simulation setting, patients were presumed to be examined until the occurrence of the outcome or event. The time of occurrence of events for each covariate was generated from a Beta $(1.25, 1)$ and further scaled on the $(0-10)$ intervals. The idea behind this interval is to distribute the measurements of the covariate across the 10-year lookback time window. We estimated the effect of treatment by restricting the lookback period correspondingly to 9 years ($t_0 - t_{-9}$), 7 years ($t_0 - t_{-7}$), five years ($t_0 - t_{-5}$), 2years ($t_0 - t_{-2}$), 1 year ($t_0 - t_{-1}$), six months ($t_0 - t_{-0.5}$), 3 months ($t_0 - t_{-0.25}$) and 1 month ($t_0 - t_{-0.083}$). Each dataset included 10,000 individuals and was replicated 1000 times for each analysis.

Following the procedure of [135], a treatment status for each individual was randomly generated from a Bernoulli distribution with subject-specific probabilities of treatment assignment. The following relation determined the selection probability for the treatment status for each subject i:

$$\begin{aligned} logit(\pi_{i,treat}) = \beta_{0,treat} + \beta_{11}c_{1t_1} + \beta_{12}c_{1t_2} + \beta_{13}c_{1t_3} + \beta_{21}c_{2t_1} + \beta_{22}c_{2t_2} \\ + \beta_{23}c_{2t_3} + \beta_{3}c_{3t_1} + \beta_{41}c_{4t_1} + \beta_{42}c_{4t_2} + \beta_{43}c_{4t_3} + \beta_{51}iv_{t_1} + \beta_{52}iv_{t_2} \end{aligned} \tag{4.1}$$

Here $\beta_{0,treat}$ indicates the treatment effect on the log-relative risk scale was determined in a preliminary simulation step. A virtual dataset of 1,000 entries was generated, and $\beta_{0,treat}$ was targeted such that the treatment would be assigned to approximately half of the individuals. The value of all other regression coefficients β_1,β_2,β_3,β_4,β_5 were kept as $log2$ to keep a moderate association. We used weight for each covariate by the time of occurrence of events using a cubic spline function with 10 degrees of freedom [136]. The term "spline" refers to a wide class of functions used in applications requiring data smoothing. A cubic spline is constructed of piece-wise third-order polynomials that pass through a set of v control points. The second derivative of each polynomial is commonly set to zero at the endpoints. A boundary condition that completes the system of v-2 equations can provide the solutions. The coefficients of the polynomials are obtained by specifying this condition. Among the variety of choices in the smoothing process, we used the choice of degree of polynomials, such as the choice of degrees of freedom and boundary conditions. Thus the above

equation is properly defined, and we specified the treatment status for each individual from a Bernoulli random variable such that: $A_i \sim Be(\pi_{i,A=1})$

For each individual, an outcome was generated conditional on treatment assignment status (A_i) following the equation:

$$\begin{aligned} log(\pi_{i,outcome}) = \beta_{0,out} + \beta_{00}A_i + \beta_{11}c_{1t_1} + \beta_{12}c_{1t_2} + \beta_{13}c_{1t_3} + \beta_{21}c_{2t_1} \\ + \beta_{22}c_{2t_2} + \beta_{23}c_{2t_3} + \beta_{3}c_{3t_1} + \beta_{41}c_{4t_1} \\ + \beta_{42}c_{4t_2} + \beta_{51}rs_{t_1} + \beta_{52}rs_{t_2} \end{aligned} \tag{4.2}$$

Finally, the outcome is defined as $Y_i \sim Bernoulli(\pi_{i,outcome})$. The value of the regression coefficients of the second equation $\beta_{00}, \beta_1, \beta_2, \beta_3, \beta_4$ and β_5 was set as $log(2)$, indicating an odds ratio of 2. A preliminary step was done to determine the value of $\beta_{0,out}$ such that the left term of the above equation was set to be negative. The idea behind this is to restrict the value of the linear predictor of the logarithm of the probability of an outcome less than 0 for all individuals and fix π_i in $(0,1)$. Thus if $Y_i = 0$, the subject would survive until the end of follow-up; otherwise, an event would occur.

4.2 Simulation Results

To estimate the ATE, we used both TMLE and IPTW methods and compared them. We implemented TMLE using a logistic regression model and SuperLearner(SL). [66]. Following the previous methodological study (Ripamonti et al, under review in Statistical Methods in Medical Research), we first examined the empirical distribution of the PS for both the treated and control group. We included the potential confounders c_1, c_2, c_3 and risk factor rs in both the propensity score model and outcome model. In the basic scenario, we considered only the first-order main effects and no interaction term/higher-order terms. First, we calculated the propensity score distribution for the entire 10years lookback period data using a logistic regression model. Then we computed the summary statistics (i.e. minimum, 1st

quantile, median, 3rd quantile, and maximum) from the estimated PS distribution.

We hypothesized these quantile values as our true parameter value θ. Similarly, for other short lookback data (9,7,5,3,2,1 years and 6,3,1 months), we estimated the PS summary statistics (θ_h) and compared them with true value. Thus we can define bias in the estimation of the propensity score as

$$Bias(\hat{\theta_h}) = E[\hat{\theta}_h - \theta]$$

Where $\hat{\theta}_h$ is the empirical maximum likelihood estimator for PS summary statistics using short lookback periods ($h = 1,2,5,7,9$ years and $6,3,1$ months). We presented the distribution of propensity scores for various lookbacks in Table 4.1. The simulation result shows that lookback 9 years produces very similar results compared to our defined reality scenario. But this situation changes surprisingly for the short term lookback (less than 5 years lookback). We can observe very poor estimates for the PS-quantiles and increased bias for lookbacks 2 years and less than 2 years compared to other lookbacks. That means a short lookback time can lead to discrepant propensity score estimation.

In Table 4.2, we estimated the average treatment effect (ATE) using PS (IPW framework; PS-IPW), TMLE-logistic, and TMLE-SL (TMLE using SuperLearner). We considered the confounders c_1, c_2, c_3, and risk factor rs in our propensity score model for the estimation of ATE. We used the inverse probability weighting method for PS-IPW in each lookback and finally estimated the average treatment effect (ATE) after the weight adjustment. We used the robust standard errors for ATE using the IPW method [82]. To do this in R, we used survey::vcovHC() after a glm() call with the outcome model recommended by [82]. We presented the marginal odds ratio for different lookbacks in Table 4.2.

For TMLE, in step 2, described in section 3.5.1, we used the same propensity score model as the PS-IPW approach. We considered the outcome model with all adjusted confounders and estimated the ATE. We present ATE for all lookbacks and corresponding standard er-

Table 4.1: Propensity Score Quantile Estimates and Bias for Varying Lookback Periods from Simulation Study

		A=0						A=1					
	10yr	Min	1st q	Med	3rd q	Max	10yr	Min	1st q	Med	3rd q	Max
		0.022	0.095	0.215	0.417	0.987		0.023	0.52	0.781	0.932	0.999
Stat	**lookback**	**Est**	**Bias** ($\hat{\theta_h}$)				**lookback**	**Est**	**Bias** ($\hat{\theta_h}$)			
Min	9yr	0.023	0.0004				9yr	0.023	0.0003			
	7yr	0.032	0.009				7yr	0.032	0.009			
	5yr	0.078	0.055				5yr	0.078	0.055			
	2yr	0.301	0.278				2yr	0.302	0.279			
	1yr	0.337	0.314				1yr	0.343	0.32			
	6mo	0.378	0.355				6mo	0.383	0.360			
	3mo	0.269	0.246				3mo	0.278	0.255			
	1mo	0.230	0.208				1mo	0.240	0.217			
1st q	9yr	0.091	0.003				9yr	0.524	0.004			
	7yr	0.078	0.016				7yr	0.534	0.014			
	5yr	0.083	0.011				5yr	0.495	0.024			
	2yr	0.309	0.214				2yr	0.31	0.21			
	1yr	0.397	0.302				1yr	0.397	0.122			
	6mo	0.434	0.339				6mo	0.434	0.085			
	3mo	0.269	0.246				3mo	0.453	0.066			
	1mo	0.230	0.208				1mo	0.461	0.058			
Med	9yr	0.209	0.005				9yr	0.786	0.005			
	7yr	0.198	0.017				7yr	0.798	0.017			
	5yr	0.218	0.003				5yr	0.78	0.0003			
	2yr	0.309	0.094				2yr	0.618	0.162			
	1yr	0.397	0.182				1yr	0.403	0.377			
	6mo	0.434	0.218				6mo	0.436	0.345			
	3mo	0.453	0.238				3mo	0.454	0.326			
	1mo	0.461	0.246				1mo	0.462	0.318			
3rd q	9yr	0.414	0.003				9yr	0.935	0.003			
	7yr	0.403	0.014				7yr	0.942	0.009			
	5yr	0.416	0.001				5yr	0.933	0.024			
	2yr	0.325	0.092				2yr	0.821	0.111			
	1yr	0.399	0.018				1yr	0.734	0.198			
	6mo	0.436	0.018				6mo	0.502	0.430			
	3mo	0.455	0.037				3mo	0.498	0.434			
	1mo	0.462	0.044				1mo	0.489	0.443			
Max	9yr	0.987	0.001				9yr	0.999	0			
	7yr	0.989	0.003				7yr	0.999	0			
	5yr	0.99	0.003				5yr	0.999	0			
	2yr	0.98	0.006				2yr	0.999	0.0005			
	1yr	0.971	0.015				1yr	0.997	0.002			
	6mo	0.964	0.022				6mo	0.991	0.008			
	3mo	0.958	0.028				3mo	0.982	0.017			
	1mo	0.952	0.034				1mo	0.963	0.036			

1) **Est** represents propensity score quantile estimates.

2) **Bias** ($\hat{\theta_h}$) represents the absolute bias

rors in Table 4.2. We also estimated the odds ratio under this approach for different lookbacks. For both PS-IPW and TMLE-logistic, no higher-order polynomials or interaction term was added. In step 2 for TMLE under SuperLearner, we used SuperLearner instead of a parametric model. In R under library SuperLearner, we used the following user-selected library for TMLE-SL ("SL.glm," "SL.step," "SL.step.interaction", "SL.glm.interaction", "SL.gam," "SL.randomForest," "SL.rpart") (see section 3.5.3). We considered the same set of covariates for the outcome model as the other two methods under TMLE-SL. The standard errors for the estimates for the TMLE and TMLE-SL were calculated based on the idea of efficient influence curve (EIC) [25] [58]. We also present the marginal odds ratio using TMLE-SL in Table 4.2. Our main results showed that the effect estimates differ slightly among the lookback periods. We assumed the true average treatment effect is 0.03 (for 10 years lookback period), and the corresponding standard error estimate is 0.0118. The effect estimates from lookback periods 9, 7, and 5 years are close to the true effect estimates. This estimate differs for smaller lookback periods (i.e., 6 months to 2 years). The estimates of standard errors also differ among the lookback periods. The lookback periods of 9 and 7 years produced estimates of standard errors close to the true value. That means, in the case of estimating ATE, longer lookback periods are more efficient and less biased compared to shorter lookbacks. We observe a similar situation for PS-IPW and TMLE-SL. For these two methods, longer lookbacks (9 and 7 years) produced ATE estimates close to the true value compared to the shorter lookbacks (1 month to 2 years).

We also observe relevant variations among the estimates of standard errors for varying lookbacks in each of the three methods. TMLE-logistic has smaller standard errors than the PS-IPW approach for lookback periods 9, 7, and 5 years. However, standard error estimates are similar in shorter lookbacks for both methods. For TMLE-SL, we noticed decreased standard errors for longer lookbacks (5, 7 and 9 years) compared to PS-IPW and TMLE-logistic. TMLE-SL also produces smaller standard error estimates for shorter lookbacks (3 months to 2 years) compared to the other two methods. So, incorporating TMLE with the SuperLearner algorithm produces less standard error estimates for longer and shorter lookback periods than TMLE-logistic and PS-IPW. The occurrence of this phenomenon can

be attributed to the SuperLearner algorithm's strategy of amalgamating numerous diverse models, which effectively captures various patterns and characteristics present in the data. It selects the optimal prediction by leveraging cross-validation [126]. In addition, for shorter lookbacks, there can be a chance of missingness in the covariate information, and choosing the appropriate model might be difficult. In this situation, TMLE with SuperLearner may provide additional opportunities for choosing the right model and producing efficient estimates, and this reflects in our simulation results.

Table 4.2: Estimates of Average Treatment Effect (ATE) and Odds Ratio (OR) Using PS-IPW, TMLE-logistic, and TMLE-SL for Varying Lookback Length

Estimation method	lookback period	ATE	SE	OR
TMLE-logistic	**10yr**	**0.0353**	**0.0118**	**1.3508**
	9yr	0.0349	0.0119	1.348
	7yr	0.0355	0.0124	1.358
	5yr	0.0432	0.0119	1.437
	2yr	0.071	0.009	1.787
	1yr	0.0821	0.0081	1.958
	6mo	0.0867	0.0077	2.0378
	3mo	0.0897	0.00746	2.09
	1mo	0.091	0.00749	2.12
PS-IPW	9yr	0.0446	0.0124	1.447
	7yr	0.0443	0.0131	1.443
	5yr	0.0504	0.0125	1.52
	2yr	0.073	0.0092	1.82
	1yr	0.0827	0.0081	1.975
	6mo	0.086	0.0077	2.045
	3mo	0.089	0.0074	2.097
	1mo	0.091	0.0072	2.122
TMLE-SL	9yr	0.0486	0.0048	1.499
	7yr	0.0451	0.005	1.458
	5yr	0.0520	0.0065	1.54
	2yr	0.0745	0.0074	1.85
	1yr	0.0787	0.0073	1.92
	6mo	0.0816	0.0073	1.97
	3mo	0.0844	0.0073	2.02
	1mo	0.0858	0.0074	2.043

Chapter 5

Analysis of Real World Data

5.1 Motivation for Real World Data analysis

The importance of assessing drug effectiveness, safety, and post-market surveillance has been widely acknowledged for over four decades. Observational studies are used to evaluate these drug assessments. Researchers from CNODES (Canadian Network for Observational Drug Effect Studies) conduct observational studies and provide extensive information on drug assessments in support of regulators and other stakeholders [7, 8, 137]. Until 2022, CNODES was a joint initiative of CIHR (Health Canada and Canadian Institutes of Health Research); it is currently funded through the Canadian Agency for Drugs and Technologies in Health. CNODES include a combined population from seven provinces (British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, and Nova Scotia) [8] as well as the CPRD (United Kingdom Clinical Practice Research Datalink), which incorporates drugs marketed in the UK [138]. In these distributed networks, the administrative data sites include minimal patient history information; thus, adopting varying lookback periods is of use in assessing confounding control. In real world evidence, having a longer lookback might not be possible as a patient's past data is less likely to be recorded due to follow up variation. However, it is very natural that treatment allocation is strongly dependent on the patient's recent characteristics but only moderately on older events. If a significant portion of recent data is missing, the patient's past data may be able to be used to compensate for this missingness. In most cases, shorter lookback time windows might capture

the most admissible information for treatment effect estimation. However, in settings such as the CPRD, patients' chronic conditions may not be recorded at each visit, and at some points, a longer lookback might be helpful. Moreover, longer lookback periods may classify patients' medications that are no longer in use, but shorter lookback may miss this medication. Choosing an optimal lookback is essential in an observational setting.

Several methods have been introduced in the literature to minimize confounding bias. However, very little work has been done to compare varying lookback periods with these approaches. Propensity score methods are commonly used to address this bias; however, these methods require the correct specification of a treatment model. However, the double robustness property of the TMLE approach can lead close to true effect estimates in the presence of model misspecification (either treatment or outcome model). As TMLE is relatively new to the literature, to our knowledge lookback periods have not been explored with this method. In this thesis, our primary motivation is to estimate the drug effects in various lookback periods (short-term and long-term) and to compare propensity score methods and TMLE. In this chapter, we use real world data for examining these issues.

5.1.1 Clinical Practice Research Datalink (CPRD)

The Clinical Practice Research Datalink (CPRD), formerly known as the "General Practice Research Database," is a UK government, non-profit primary care research database that collects anonymized patient data from 624 general practices across the UK, covering 6.3% of all 9949 practices in 2012 [139]. In October 2017, CPRD initiated a new CPRD AURUM database that collects data routinely from practices using electronic patient record (EHR) system software [140]. As of September 2018, the CPRD database contains 7 million patients portraying 13% of the population in England. So, this significant population-based data source CPRD AURUM has extended historical information on individuals from various geographical regions linked to secondary care, disease registries, and death registration records. We use this database as an ideal to create different lookback data and to conduct our research.

5.1.2 Description of the Data

Study Population

The study cohort was drawn from patients receiving their first-ever prescriptions for either metformin or sulfonylureas between April 1, 1998, and December 30, 2019. Patients meeting the following criteria were excluded: 1) age < 18 years; 2) >= 1 year of database history. 3) A previous prescription for an antidiabetic drug; 4) A prescription for both metformin and sulfonylureas or a prescription for any antidiabetic drug other than metformin or a sulfonylurea on the day of cohort entry; 5) A previous diagnosis of polycystic ovary syndrome (other indication for metformin use); 6) No recorded follow-up. Patients were followed until an event (defined below) occurred, a departure from the CPRD or HES (Hospital Episode Statistics) , or the end of the study period (December 30, 2019), whichever occurs first.

Exposure Definition

Exposure was defined as the initial use of sulfonylurea monotherapy. Switching within sulfonylureas (i.e., glyburide, glimepiride, and glipizide) is permissible, but patients were not allowed to change their exposure (i.e., metformin to sulfonylurea or vice-versa) as it might be an indicator of diabetes severity or could be related to an adverse drug event. Exposure was treated as **time-fixed** in a preliminary analysis, and an **intention-to-treat** approach will be used. Patients were considered exposed to either drug following their first medication regardless of the treatment switches or discontinuation. To minimize exposure misclassification, the follow-up period was restricted to a maximum of 1 year.

Outcome Definition

The outcome of interest for this methodological study was death due to major adverse cardiovascular events during the first year of post-treatment initiation. We have assessed the outcome based on recorded diagnoses in CPRD AURUM, HES. We identified HES records with the corresponding ICD-10 codes in the primary or secondary position or as the underlying cause of death (for myocardial infarction, 121. X; for ischemic stroke, 163. X, 164. X). HES data was supplemented by ONS (Office of National Statistics) data. Cardiovascular

death (ICD-10 codes 10-99) was defined using ONS data, and all-cause mortality were determined using both data sources and CPRD AURUM. The date of death was used as the event date.

Covariates

A range of covariates were adjusted in our TMLE and PS-IPW analyses. These covariates measure in the year before cohort entry include demographic characteristics (age, sex, calendar year), socio-economic status, lifestyle variables (smoking [ever, never], excessive alcohol use), BMI, previous medical history (blood pressure, atrial fibrillation or flutter, cancer, cerebrovascular disease, chronic obstructive pulmonary disease, coronary heart disease, heart failure, hyperlipidemia, hypertension, previous myocardial infarction, previous coronary revascularization, previous stroke, and thyroid disease), complications of diabetes (neuropathy, peripheral arterial or vascular disease, renal disease, and retinal disorder), eGFR, and glycated hemoglobin level ($7\%, 7.1-8.0\%, > 8\%$, unknown). We also assessed the use of the following medications: angiotensin-converting enzyme inhibitors, angiotensin II receptor blockers, diuretics, digoxin, statins, fibrates, acetylsalicylic acid, clopidogrel, warfarin, nonsteroidal, anti-inflammatory drugs, opioid analgesics, and paracetamol. To evaluate the effect of measurement error due to varying lookback periods on using PS-IPW and TMLE, we evaluated covariates (comorbidities, drugs, lifestyle variables, number of hospitalizations) with a lookback of 10 years and then restrict the lookback to 1, 3, 6 months, and 1, 2, 5, 7 years.

5.1.3 Analysis and Results

This is a retrospective population-based cohort study with a total number of 656,450 study participants. We extracted a cohort of eligible study participants for each lookback period (1,3,6 months and 1,2,5,7,10 years). In each lookback, a total number of 556,667 metformin and 99,783 sulfonylurea users were included. Study participants in the metformin group consist of 301,070 males and 255,597 females, and for the sulfonylureas group, 42,751 females and 57,032 males. For both groups, the age of most participants is more than 50

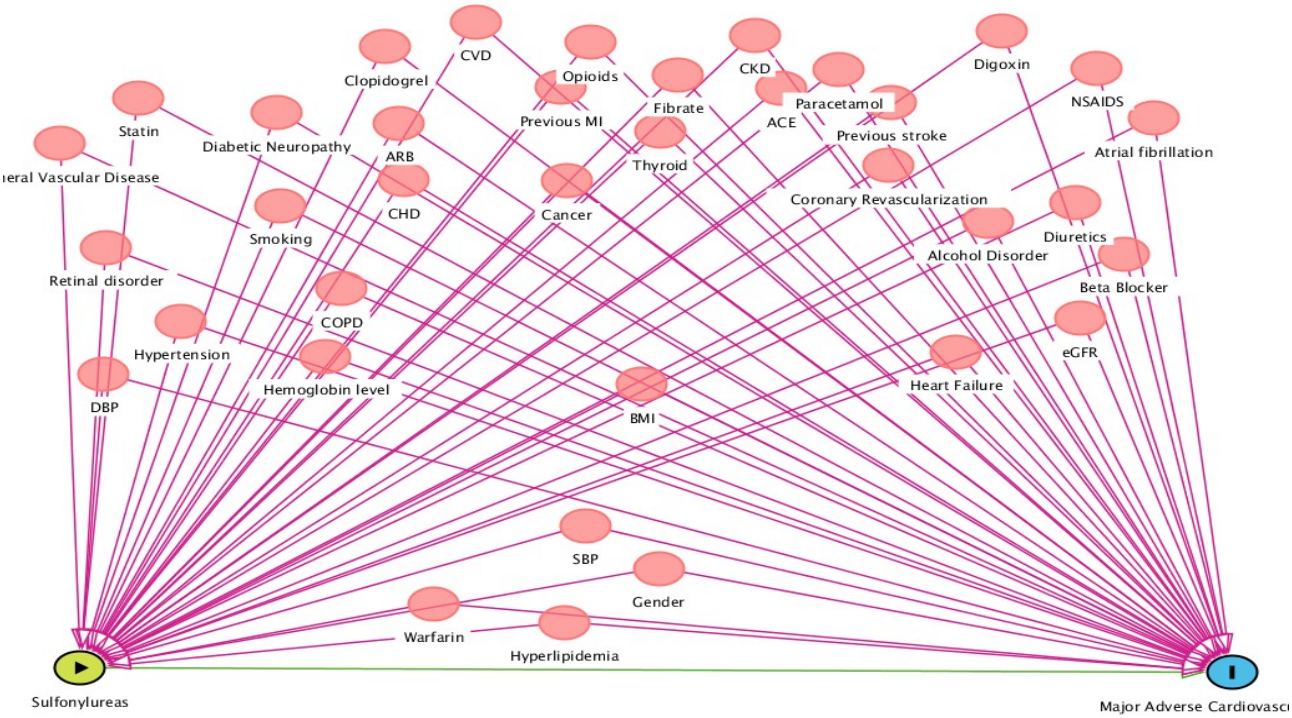


Figure 5.1: Causal diagram with confounding variables

years (70% for the metformin and 83.4% for sulfonylureas) in all lookback periods. For each lookback, the percentage of the smoking group is higher than non-smoking (Table 5.1, 5.2). The patient's other demographic and lifestyle characteristics, comorbidities and risk factors, and medication use for each lookback are presented in Table 5.1, 5.2, and Table 5.3.

Table 5.1: Demographical and Clinical Characteristics of Study Participants (Lookback Period 1, 3 and 6 months)

Variables	lookback period 1mo		lookback Period 3mo		lookback Period 6mo	
	Metformin (N=556,667)	Sulfonylureas (N=99,783)	Metformin (N=556,667)	Sulfonylureas (N=99,783)	Metformin N=556,667	Sulfonylureas (N=99,783)
Gender						
Female	255597(45.9)	42751(42.8)	255597(45.9)	42751(42.8)	255597(45.9)	42751(42.8)
Male	301070(54.1)	57032(57.2)	301070(54.1)	57032(57.2)	301070(54.1)	57032(57.2)
Age						
Less than 20	1586 (0.3)	93 (0.1)	1586 (0.3)	93 (0.1)	1586 (0.3)	93 (0.1)
20 to 30	16624 (3.0)	1022 (1.0)	16624 (3.0)	1022 (1.0)	16624 (3.0)	1022 (1.0)
30 to 40	42092 (7.6)	4486 (4.5)	42092 (7.6)	4486 (4.5)	42092 (7.6)	4486 (4.5)
40 to 50	92800 (16.7)	10913(10.9)	92800 (16.7)	10913(10.9)	92800 (16.7)	10913(10.9)
50 to 60	138438(24.9)	18861(18.9)	138438(24.9)	18861(18.9)	138438(24.9)	18861(18.9)
60 to 70	140261(25.2)	25973(26.0)	140261(25.2)	25973(26.0)	140261(25.2)	25973(26.0)
>70	124866(22.4)	38435(38.5)	124866(22.4)	38435(38.5)	124866(22.4)	38435(38.5)
Calendar year entry						
1998-2003	73517 (13.2)	60003(60.1)	73517 (13.2)	60003(60.1)	73517 (13.2)	60003(60.1)
2004-2009	163636 (29.4)	21960(22.0)	163636(29.4)	21960(22.0)	163636(29.4)	21960(22.0)
2010-2014	154100 (27.7)	10296(10.3)	154100(27.7)	10296(10.3)	154100(27.7)	10296(10.3)
2015-2019	165414 (29.7)	7524 (7.5)	165414(29.7)	7524 (7.5)	165414(29.7)	7524 (7.5)
Comorbidities and risk factors						
Smoking status						
Never smoking	95673 (17.2)	9768 (9.8)	119810 (21.5)	13167 (13.2)	140840 (25.3)	140840 (25.3)
Ever smoking	136478 (24.5)	12645 (12.7)	175239 (31.5)	17620 (17.7)	209902 (37.7)	22403 (22.5)
Missing	324516 (58.3)	77370 (77.5)	261618 (47.0)	68996 (69.1)	205925 (37.0)	61025 (61.2)
Alcohol disorder						
No	541796 (97.3)	98136 (98.3)	537107 (96.5)	97469 (97.7)	532360 (95.6)	96799 (97.0)
Yes	14871 (2.7)	1647 (1.7)	19560 (3.5)	2314 (2.3)	24307 (4.4)	2984 (3.0)
BMI						
<25	21586 (3.9)	10261 (10.3)	26194 (4.7)	12553 (12.6)	30235 (5.4)	14318 (14.3)
25-30	75855 (13.6)	11073 (11.1)	91100 (16.4)	14222 (14.3)	103734 (18.6)	16764 (16.8)
>30	160992 (28.9)	7354 (7.4)	193372 (34.7)	9680 (9.7)	218587 (39.3)	11794 (11.8)
Missing	298234 (53.6)	71095 (71.2)	246001 (44.2)	63328 (63.5)	204111 (36.7)	56907(57.0)
Hemoglobin level						
<7	71786 (12.9)	2917 (2.9)	91201 (16.4)	4456 (4.5)	99067 (17.8)	5883 (5.9)
7-8	109039 (19.6)	7537 (7.6)	131134 (23.6)	9724 (9.7)	136724 (24.6)	10896 (10.9)
>8	154051 (27.7)	23991 (24.0)	175583 (31.5)	28308 (28.4)	179752 (32.3)	29686 (29.8)
Missing	221791 (39.8)	65338 (65.5)	158749 (28.5)	57295 (57.4)	141124 (25.4)	53318 (53.4)
Systolic blood pressure						
Mean (SD)	138 (17.5)	140 (20.7)	137 (17.2)	140 (20.5)	137 (17.1)	140 (20.3)
Missing	221791 (39.8)	65338 (65.5)	140629 (25.3)	43089 (43.2)	95548 (17.2)	34386 (34.5)
Diastolic blood pressure						
Mean (SD)	80.9 (10.7)	80.2 (11.4)	80.6 (10.6)	80.0 (11.3)	80.4 (10.5)	79.9 (11.2)
Missing	211769 (38.0)	56070 (56.2)	140629 (25.3)	43089 (43.2)	95548 (17.2)	34386 (34.5)
Atrial fibrillation						
No	555888 (99.9)	99606 (99.8)	555268 (99.7)	99472 (99.7)	554504 (99.6)	99307 (99.5)
Yes	779 (0.1)	177 (0.2)	1399 (0.3)	311 (0.3)	2163 (0.4)	476 (0.5)
Cancer						
No	553581 (99.4)	97828 (98.0)	551267 (99.0)	96744 (97.0)	549061 (98.6)	95912 (96.1)
Yes	3086 (0.6)	1955 (2.0)	5400 (1.0)	3039 (3.0)	7606 (1.4)	3871 (3.9)
Cerebrovascular disease						
No	552255 (99.2)	98704 (98.9)	549504 (98.7)	98027 (98.2)	547153 (98.3)	97371 (97.6)
Yes	4412 (0.8)	1079 (1.1)	7163 (1.3)	1756 (1.8)	9514 (1.7)	2412 (2.4)

All values are presented as n (%), unless otherwise stated.

Variables	lookback period 1mo		lookback period 3mo		lookback period 6mo	
	Metformin	Sulfonylureas	Metformin	Sulfonylureas	Metformin	Sulfonylureas
Chronic Obstructive Pulmonary Disease						
No	551732 (99.1)	99000 (99.2)	548193 (98.5)	98431 (98.6)	544914 (97.9)	97892 (98.1)
Yes	4935 (0.9)	783 (0.8)	8474 (1.5)	1352 (1.4)	11753 (2.1)	1891 (1.9)
Coronary heart disease						
No	539832 (97.0)	96337 (96.5)	530938 (95.4)	94504 (94.7)	522408 (93.8)	92719 (92.9)
Yes	16835 (3.0)	3446 (3.5)	25729 (4.6)	5279 (5.3)	34259 (6.2)	7064 (7.1)
Heart failure						
No	554690 (99.6)	98857 (99.1)	553324 (99.4)	98266 (98.5)	551958 (99.2)	97693 (97.9)
Yes	1977 (0.4)	926 (0.9)	3343 (0.6)	1517 (1.5)	4709 (0.8)	2090 (2.1)
Hyperlipidemia						
No	541082 (97.2)	97803 (98.0)	535522 (96.2)	96884 (97.1)	529771 (95.2)	95923 (96.1)
Yes	15585 (2.8)	1980 (2.0)	21145 (3.8)	2899 (2.9)	26896 (4.8)	3860 (3.9)
Hypertension						
No	499809 (89.8)	93306 (93.5)	471484 (84.7)	89599 (89.8)	446884 (80.3)	85959 (86.1)
Yes	56858 (10.2)	6477 (6.5)	85183 (15.3)	10184 (10.2)	109783 (19.7)	13824 (13.9)
Coronary Revascularization						
No	556054 (99.9)	99507 (99.7)	555560 (99.8)	99370 (99.6)	554976 (99.7)	99265 (99.5)
Yes	613 (0.1)	276 (0.3)	1107 (0.2)	413 (0.4)	1691 (0.3)	518 (0.5)
Thyroid						
No	551168 (99.0)	99132 (99.3)	548454 (98.5)	98750 (99.0)	545778 (98.0)	98352 (98.6)
Yes	5499 (1.0)	651 (0.7)	8213 (1.5)	1033 (1.0)	10889 (2.0)	1431 (1.4)
Diabetic Neuropathy						
No	545239 (97.9)	98847 (99.1)	542352 (97.4)	98510 (98.7)	538642 (96.8)	98122 (98.3)
Yes	11428 (2.1)	936 (0.9)	14315 (2.6)	1273 (1.3)	18025 (3.2)	1661 (1.7)
Peripheral Vascular Disease						
No	554699 (99.6)	99332 (99.5)	553679 (99.5)	99063 (99.3)	552659 (99.3)	98779 (99.0)
Yes	1968 (0.4)	451 (0.5)	2988 (0.5)	720 (0.7)	4008 (0.7)	1004 (1.0)
Chronic Kidney disease						
No	551930 (99.1)	98705 (98.9)	549538 (98.7)	98107 (98.3)	547039 (98.3)	97459 (97.7)
Yes	4737 (0.9)	1078 (1.1)	7129 (1.3)	1676 (1.7)	9628 (1.7)	2324 (2.3)
Retinal disorder						
No	553828 (99.5)	99385 (99.6)	550798 (98.9)	99065 (99.3)	547638 (98.4)	98693 (98.9)
Yes	2839 (0.5)	398 (0.4)	5869 (1.1)	718 (0.7)	9029 (1.6)	1090 (1.1)
eGFR						
Mean (SD)	81.1 (18.9)	69.8 (23.6)	81.1 (18.8)	69.4 (23.5)	81.1 (18.8)	69.3 (23.3)
Missing	272200 (48.9)	65237 (65.4)	179984 (32.3)	54412 (54.5)	134114 (24.1)	47663 (47.8)
Previous stroke						
No	553937 (99.5)	99092 (99.3)	552054 (99.2)	98676 (98.9)	550540 (98.9)	98294 (98.5)
Yes	2730 (0.5)	691 (0.7)	4613 (0.8)	1107 (1.1)	6127 (1.1)	6127 (1.1)
Previous Myocardial Infarction						
No	554609 (99.6)	98865 (99.1)	553501 (99.4)	98472 (98.7)	552422 (99.2)	98171 (98.4)
Yes	2058 (0.4)	918 (0.9)	3166 (0.6)	1311 (1.3)	4245 (0.8)	1612 (1.6)
MajorAdverse Cardiovascular Events (MACE)						
No	467178 (83.9)	56843 (57.0)	467178 (83.9)	56843 (57.0)	467178 (83.9)	56843 (57.0)
Yes	89489 (16.1)	42940(43.0)	89489 (16.1)	42940(43.0)	89489 (16.1)	42940(43.0)

All values are presented as n (%), unless otherwise stated.

Variables	lookback Period 1mo		lookback Period 3mo		lookback Period 6mo	
	Metformin	Sulfonylureas	Metformin	Sulfonylureas	Metformin	Sulfonylureas
Medications						
Angiotensin Converting Enzyme (ACE)						
No	423659 (76.1)	80648 (80.8)	392108 (70.4)	76238 (76.4)	387113 (69.5)	75045 (75.2)
Yes	133008 (23.9)	19135 (19.2)	164559 (29.6)	23545 (23.6)	169554 (30.5)	24738 (24.8)
Angiotensin Receptor Blockers (ARB)						
No	512317 (92.0)	95187 (95.4)	500205 (89.9)	93922 (94.1)	498663 (89.6)	93569 (93.8)
Yes	44350 (8.0)	4596 (4.6)	56462 (10.1)	5861 (5.9)	58004 (10.4)	6214 (6.2)
Beta Blocker						
No	467124 (83.9)	82825 (83.0)	441020 (79.2)	77812 (78.0)	436770 (78.5)	76603 (76.8)
Yes	89543 (16.1)	16958 (17.0)	115647 (20.8)	21971 (22.0)	119897 (21.5)	23180 (23.2)
Diuretics						
No	454717 (81.7)	76946 (77.1)	422741 (75.9)	70288 (70.4)	416046 (74.7)	68472 (68.6)
Yes	101950 (18.3)	22837 (22.9)	133926 (24.1)	29495 (29.6)	140621 (25.3)	31311 (31.4)
Digoxin						
No	545234 (97.9)	94288 (94.5)	542518 (97.5)	92956 (93.2)	542208 (97.4)	92741 (92.9)
Yes	11433 (2.1)	5495 (5.5)	14149 (2.5)	6827 (6.8)	14459 (2.6)	7042 (7.1)
Statin						
No	328327 (59.0)	77288 (77.5)	280920 (50.5)	72398 (72.6)	274521 (49.3)	71433 (71.6)
Yes	228340 (41.0)	22495 (22.5)	275747 (49.5)	27385 (27.4)	282146 (50.7)	28350 (28.4)
Nonsteroidal Anti inflammatory						
Drugs (NSAIDS)						
No	509432 (91.5)	92151 (92.4)	478544 (86.0)	87075 (87.3)	451411 (81.1)	82655 (82.8)
Yes	47235 (8.5)	7632 (7.6)	78123 (14.0)	12708 (12.7)	105256 (18.9)	17128 (17.2)
Opioids						
No	482717 (86.7)	85679 (85.9)	450554 (80.9)	79553 (79.7)	426638 (76.6)	75231 (75.4)
Yes	73950 (13.3)	14104 (14.1)	106113 (19.1)	20230 (20.3)	130029 (23.4)	24552 (24.6)
Paracetamol						
No	477541 (85.8)	83919 (84.1)	440201 (79.1)	76726 (76.9)	415356 (74.6)	72007 (72.2)
yes	79126 (14.2)	15864 (15.9)	116466 (20.9)	23057 (23.1)	141311 (25.4)	27776 (27.8)
Fibrate						
No	551632 (99.1)	98928 (99.1)	550209 (98.8)	98685 (98.9)	549804 (98.8)	98569 (98.8)
yes	5035 (0.9)	855 (0.9)	6458 (1.2)	1098 (1.1)	6863 (1.2)	1214 (1.2)
Aspirin						
No	465169 (83.6)	81769 (81.9)	439325 (78.9)	76532 (76.7)	432375 (77.7)	74721 (74.9)
yes	91498 (16.4)	18014 (18.1)	117342 (21.1)	23251 (23.3)	124292 (22.3)	25062 (25.1)
Clopidogrel						
No	544294 (97.8)	97774 (98.0)	541273 (97.2)	97386 (97.6)	540178 (97.0)	97232 (97.4)
yes	12373 (2.2)	2009 (2.0)	15394 (2.8)	2397 (2.4)	16489 (3.0)	2551 (2.6)
Warfarin						
No	542728 (97.5)	96132 (96.3)	537424 (96.5)	94645 (94.9)	536164 (96.3)	94204 (94.4)
yes	13939 (2.5)	3651 (3.7)	19243 (3.5)	5138 (5.1)	20503 (3.7)	5579 (5.6)

All values are presented as n (%), unless otherwise stated.

Table 5.2: Demographical and Clinical Characteristics of Study Participants (Lookback Period 1, 2 and 5 years

Variables	lookback period 1yr		lookback Period 2yr		lookback Period 5yr	
	Metformin (N=556,667)	Sulfonylureas (N=99,783)	Metformin (N=556,667)	Sulfonylureas (N=99,783)	Metformin N=556,667	Sulfonylureas (N=99,783)
Gender						
Female	255597(45.9)	42751(42.8)	255597(45.9)	42751(42.8)	255597(45.9)	42751(42.8)
Male	301070(54.1)	57032(57.2)	301070(54.1)	57032(57.2)	301070(54.1)	57032(57.2)
Age						
Less than 20	1586 (0.3)	93 (0.1)	1586 (0.3)	93 (0.1)	1586 (0.3)	93 (0.1)
20 to 30	16624 (3.0)	1022 (1.0)	16624 (3.0)	1022 (1.0)	16624 (3.0)	1022 (1.0)
30 to 40	42092 (7.6)	4486 (4.5)	42092 (7.6)	4486 (4.5)	42092 (7.6)	4486 (4.5)
40 to 50	92800 (16.7)	10913(10.9)	92800 (16.7)	10913(10.9)	92800 (16.7)	10913(10.9)
50 to 60	138438(24.9)	18861(18.9)	138438(24.9)	18861(18.9)	138438(24.9)	18861(18.9)
60 to 70	140261(25.2)	25973(26.0)	140261(25.2)	25973(26.0)	140261(25.2)	25973(26.0)
>70	124866(22.4)	38435(38.5)	124866(22.4)	38435(38.5)	124866(22.4)	38435(38.5)
Calendar year entry						
1998-2003	73517 (13.2)	60003(60.1)	73517 (13.2)	60003(60.1)	73517 (13.2)	60003(60.1)
2004-2009	163636 (29.4)	21960(22.0)	163636(29.4)	21960(22.0)	163636(29.4)	21960(22.0)
2010-2014	154100 (27.7)	10296(10.3)	154100(27.7)	10296(10.3)	154100(27.7)	10296(10.3)
2015-2019	165414 (29.7)	7524 (7.5)	165414(29.7)	7524 (7.5)	165414(29.7)	7524 (7.5)
Comorbidities and risk factors						
Smoking status						
Never smoking	164957 (29.6)	20209(20.3)	181698(32.6)	23843(23.9)	189450(34.0)	28326(28.4)
Ever smoking	250481 (45.0)	28386(28.4)	286058(51.4)	34638(34.7)	324317(58.3)	43090(43.2)
Missing	141229 (25.4)	51188(51.3)	88911 (16.0)	41302(41.4)	42900 (7.7)	28367(28.4)
Alcohol disorder						
No	524750 (94.3)	95846(96.1)	511682(91.9)	94428(94.6)	487127(87.5)	91744(91.9)
Yes	31917 (5.7)	3937 (3.9)	44985 (8.1)	5355 (5.4)	69540 (12.5)	8039 (8.1)
BMI						
<25	35060 (6.3)	16313(16.3)	39944 (7.2)	18176(18.2)	45894 (8.2)	20676(20.7)
25-30	118060 (21.2)	19725(19.8)	130743(23.5)	22569(22.6)	144875(26.0)	26831(26.9)
>30	247753 (44.5)	14281(14.3)	273239(49.1)	16798(16.8)	298809(53.7)	20579(20.6)
Missing	155794 (28.0)	49464(49.6)	112741(20.3)	42240(42.3)	67089 (12.1)	31697(31.8)
Hemoglobin level						
<7	106309 (19.1)	7648 (7.7)	112144(20.1)	8871 (8.9)	11737(21.1)	9894 (9.9)
7-8	139538 (25.1)	11856(11.9)	140792(25.3)	12334(12.4)	141315(25.4)	12542(12.6)
>8	181635 (32.6)	30489(30.6)	182505(32.8)	30927(31.0)	183119(32.9)	31190(31.3)
Missing	129185 (23.2)	49790(49.9)	121226(21.8)	47651(47.8)	114859(20.6)	46157(46.3)
Systolic blood pressure						
Mean (SD)	137 (17.0)	140 (20.2)	137 (17.1)	140 (20.3)	136 (17.1)	140 (20.3)
Missing	60607 (10.9)	26708(26.8)	38730 (7.0)	21072(21.1)	19343 (3.5)	14433(14.5)
Diastolic blood pressure						
Mean (SD)	80.3 (10.4)	79.9 (11.1)	80.3 (10.4)	80.0 (11.1)	80.3 (10.4)	80.2 (11.1)
Missing	60607 (10.9)	26708(26.8)	38730 (7.0)	21072(21.1)	19343 (3.5)	14433(14.5)
Atrial fibrillation						
No	553179(99.4)	99035(99.3)	551017(99.0)	98674(98.9)	547117(98.3)	98075(98.3)
Yes	3488 (0.6)	748 (0.7)	5650 (1.0)	1109 (1.1)	9550 (1.7)	1708 (1.7)
Cancer						
No	545779 (98.0)	95030(95.2)	541043(97.2)	93876(94.1)	532553(95.7)	92125(92.3)
Yes	10888 (2.0)	4753 (4.8)	15624 (2.8)	5907 (5.9)	24114 (4.3)	7658 (7.7)
Cerebrovascular disease						
No	543805 (97.7)	96587(96.8)	539107(96.8)	95471(95.7)	531234(95.4)	93627(93.8)
Yes	12862 (2.3)	3196 (3.2)	17560 (3.2)	4312 (4.3)	25433 (4.6)	6156 (6.2)

All values are presented as n (%), unless otherwise stated.

Variables	lookback period 1yr		lookback period 2yr		lookback period 5yr	
	Metformin	Sulfonylureas	Metformin	Sulfonylureas	Metformin	Sulfonylureas
Chronic Obstructive Pulmonary Disease						
No	540518 (97.1)	97223(97.4)	536282(96.3)	96480(96.7)	531746(95.5)	95530(95.7)
Yes	16149 (2.9)	2560 (2.6)	20385 (3.7)	3303 (3.3)	24921 (4.5)	4253 (4.3)
Coronary heart disease						
No	510758 (91.8)	90275(90.5)	497824(89.4)	87390(87.6)	479559(86.1)	83196(83.4)
Yes	45909 (8.2)	9508 (9.5)	58843 (10.6)	12393(12.4)	77108 (13.9)	16587(16.6)
Heart failure						
No	549857 (98.8)	96966(97.2)	547093(98.3)	96055(96.3)	542542(97.5)	94528(94.7)
Yes	6810 (1.2)	2817 (2.8)	9574 (1.7)	3728 (3.7)	14125 (2.5)	5255 (5.3)
Hyperlipidemia						
No	521018 (93.6)	94550(94.8)	506573(91.0)	92582(92.8)	476382(85.6)	88994(89.2)
Yes	35649 (6.4)	5233 (5.2)	50094 (9.0)	7201 (7.2)	80285 (14.4)	10789(10.8)
Hypertension						
No	413513 (74.3)	81162(81.3)	373588(67.1)	75275(75.4)	322561(57.9)	67048(67.2)
Yes	143154 (25.7)	18621(18.7)	183079(32.9)	24508(24.6)	234106(42.1)	32735(32.8)
Coronary Revascularization						
No	553948 (99.5)	99081(99.3)	552060(99.2)	98755(99.0)	547572(98.4)	98017(98.2)
Yes	2719 (0.5)	702 (0.7)	4607 (0.8)	1028 (1.0)	9095 (1.6)	1766 (1.8)
Thyroid						
No	541689 (97.3)	97813(98.0)	535768(96.2)	96934(97.1)	524847(94.3)	95476(95.7)
Yes	14978 (2.7)	1970 (2.0)	20899 (3.8)	2849 (2.9)	31820 (5.7)	4307 (4.3)
Diabetic Neuropathy						
No	532972 (95.7)	97560(97.8)	528107(94.9)	97036(97.2)	523004(94.0)	96620(96.8)
Yes	23695 (4.3)	2223 (2.2)	28560 (5.1)	2747 (2.8)	33663 (6.0)	3163 (3.2)
Peripheral Vascular Disease						
No	551039 (99.0)	98350(98.6)	548794(98.6)	97788(98.0)	544795(97.9)	96806(97.0)
Yes	5628 (1.0)	1433 (1.4)	7873 (1.4)	1995 (2.0)	11872 (2.1)	2977 (3.0)
Chronic kidney disease						
No	542788 (97.5)	96367(96.6)	536416(96.4)	95046(95.3)	525751(94.4)	93241(93.4)
Yes	13879 (2.5)	3416 (3.4)	20251 (3.6)	4737 (4.7)	30916 (5.6)	6542 (6.6)
Retinal disorder						
No	542428 (97.4)	98169(98.4)	536931(96.5)	97500(97.7)	530330(95.3)	96707(96.9)
Yes	14239 (2.6)	1614 (1.6)	19736 (3.5)	2283 (2.3)	26337 (4.7)	3076 (3.1)
eGFR						
Mean (SD)	81.2 (18.8)	69.2 (23.1)	81.4 (18.9)	69.3 (23.1)	81.7 (18.9)	69.4 (23.1)
Missing	97847 (17.6)	41801(41.9)	79742 (14.3)	38363(38.4)	64422 (11.6)	35852(35.9)
Previous stroke						
No	548365 (98.5)	97835(98.0)	545173(97.9)	97171(97.4)	539601(96.9)	95993(96.2)
Yes	8302 (1.5)	1948 (2.0)	11494 (2.1)	2612 (2.6)	17066 (3.1)	3790 (3.8)
Previous Myocardial Infarction						
No	550793 (98.9)	97751(98.0)	547933(98.4)	97103(97.3)	541697(97.3)	95564(95.8)
Yes	5874 (1.1)	2032 (2.0)	8734 (1.6)	2680 (2.7)	14970 (2.7)	4219 (4.2)
Major Adverse Cardiovascular Events						
No	467178 (83.9)	56843(57.0)	467178(83.9)	56843(57.0)	467178(83.9)	56843(57.0)
Yes	89489 (16.1)	42940(43.0)	89489 (16.1)	42940(43.0)	89489 (16.1)	42940(43.0)

All values are presented as n (%), unless otherwise stated.

Variables	lookback Period 1yr		lookback Period 2yr		lookback Period 5yr	
	Metformin	Sulfonylureas	Metformin	Sulfonylureas	Metformin	Sulfonylureas
Medications						
Angiotensin Converting Enzyme (ACE)						
No	381059 (68.5)	73699(73.9)	370980(66.6)	71873(72.0)	350846(63.0)	68999(69.1)
Yes	175608 (31.5)	26084(26.1)	185687(33.4)	27910(28.0)	205821(37.0)	30784(30.9)
Angiotensin Receptor Blockers (ARB)						
No	497003 (89.3)	93213(93.4)	494431(88.8)	92729(92.9)	489439(87.9)	92086(92.3)
Yes	59664 (10.7)	6570 (6.6)	62236 (11.2)	7054 (7.1)	67228 (12.1)	7697 (7.7)
Beta Blocker						
No	431827 (77.6)	75393(75.6)	423599(76.1)	73690(73.9)	404092(72.6)	70206(70.4)
Yes	124840 (22.4)	24390(24.4)	133068(23.9)	26093(26.1)	152575(27.4)	29577(29.6)
Diuretics						
No	408794 (73.4)	66900(67.0)	397661(71.4)	64823(65.0)	375331(67.4)	61352(61.5)
Yes	147873 (26.6)	32883(33.0)	159006(28.6)	34960(35.0)	181336(32.6)	38431(38.5)
Digoxin						
No	541962 (97.4)	92618(92.8)	541559(97.3)	92451(92.7)	540644(97.1)	92150(92.4)
Yes	14705 (2.6)	7165 (7.2)	15108 (2.7)	7332 (7.3)	16023 (2.9)	7633 (7.6)
Statin						
No	269856 (48.5)	70726(70.9)	264247(47.5)	69943(70.1)	255778(45.9)	68960(69.1)
Yes	286811 (51.5)	29057(29.1)	292420(52.5)	29840(29.9)	300889(54.1)	30823(30.9)
Nonsteroidal Anti inflammatory Drugs (NSAIDS)						
No	410190 (73.7)	76544(76.7)	351916(63.2)	68069(68.2)	258783(46.5)	55047(55.2)
Yes	146477 (26.3)	23239(23.3)	204751(36.8)	31714(31.8)	297884(53.5)	44736(44.8)
Opioids						
No	393322 (70.7)	69844(70.0)	347796(62.5)	63087(63.2)	276983(49.8)	53593(53.7)
Yes	163345 (29.3)	29939(30.0)	208871(37.5)	36696(36.8)	279684(50.2)	46190(46.3)
Paracetamol						
No	384048 (69.0)	66619(66.8)	342534(61.5)	60420(60.6)	277378(49.8)	51891(52.0)
yes	172619 (31.0)	33164(33.2)	214133(38.5)	39363(39.4)	279289(50.2)	47892(48.0)
Fibrate						
No	549236 (98.7)	98420 (98.6)	548231 (98.5)	98183 (98.4)	545828 (98.1)	97698 (97.9)
yes	7431 (1.3)	1363 (1.4)	8436 (1.5)	1600 (1.6)	10839 (1.9)	2085 (2.1)
Aspirin						
No	425968 (76.5)	73360 (73.5)	417030 (74.9)	71728 (71.9)	399951 (71.8)	69167 (69.3)
yes	130699 (23.5)	26423 (26.5)	139637 (25.1)	28055 (28.1)		
clopidogrel						
No	538550 (96.7)	97001 (97.2)	535840 (96.3)	96689 (96.9)	530286 (95.3)	96197 (96.4)
yes	18117 (3.3)	2782 (2.8)	20827 (3.7)	3094 (3.1)	26381 (4.7)	3586 (3.6)
Warfarin						
No	535126 (96.1)	93881 (94.1)	533514 (95.8)	93505 (93.7)	530163 (95.2)	92875 (93.1)
yes	21541 (3.9)	5902 (5.9)	23153 (4.2)	6278 (6.3)	26504 (4.8)	6908 (6.9)

All values are presented as n (%), unless otherwise stated.

Table 5.3: Demographical and Clinical Characteristics of Study Participants (Lookback Period 7 and 10 years)

Variables	lookback period 7yr		lookback period 10yr	
	Metformin	Sulfonylureas	Metformin	Sulfonylureas
Gender				
Female	255597(45.9)	42751(42.8)	255597(45.9)	42751(42.8)
Male	301070(54.1)	57032(57.2)	301070(54.1)	57032(57.2)
Age				
Less than 20	1586 (0.3)	93 (0.1)	1586 (0.3)	93 (0.1)
20 to 30	16624 (3.0)	1022 (1.0)	16624 (3.0)	1022 (1.0)
30 to 40	42092 (7.6)	4486 (4.5)	42092 (7.6)	4486 (4.5)
40 to 50	92800 (16.7)	10913(10.9)	92800(16.7)	10913(10.9)
50 to 60	138438(24.9)	18861(18.9)	138438(24.9)	18861(18.9)
60 to 70	140261(25.2)	25973(26.0)	140261(25.2)	25973(26.0)
>70	124866(22.4)	38435(38.5)	124866(22.4)	38435(38.5)
Calendar year entry				
1998-2003	73517 (13.2)	60003(60.1)	73517 (13.2)	60003(60.1)
2004-2009	163636(29.4)	21960(22.0)	163636(29.4)	21960(22.0)
2010-2014	154100(27.7)	10296(10.3)	154100(27.7)	10296(10.3)
2015-2019	165414(29.7)	7524 (7.5)	165414(29.7)	7524 (7.5)
Comorbidities and risk factors				
BMI				
<25	47784 (8.6)	21865(21.9)	49541 (8.9)	22937(23.0)
25-30	149381(26.8)	29145(29.2)	153423(27.6)	30957(31.0)
>30	305825(54.9)	22281(22.3)	311012(55.9)	23375(23.4)
Missing	53677 (9.6)	26492(26.5)	42691 (7.7)	22514(22.6)
Smoking status				
Never smoking	188383(33.8)	30546(30.6)	185419(33.3)	31872(31.9)
Ever smoking	337745(60.7)	47213(47.3)	350425(63.0)	50654(50.8)
Missing	30539 (5.5)	22024(22.1)	20823(3.7)	17257(17.3)
Alcohol disorder				
No	476500(85.6)	90571(90.8)	466431(83.8)	89495(89.7)
Yes	80167 (14.4)	9212 (9.2)	90236 (16.2)	10288(10.3)
Hemoglobin level				
<7	118516(21.3)	10092(10.1)	119404(21.4)	10268(10.3)
7-8	141381(25.4)	12558(12.6)	141436(25.4)	12598(12.6)
>8	183315(32.9)	31236(31.3)	183626(33.0)	31302(31.4)
Missing	113455(20.4)	45897(46.0)	112201(20.2)	45615(45.7)
Systolic blood pressure				
Mean (SD)	136 (17.1)	140 (20.3)	136 (17.1)	140 (20.3)
Missing	15386 (2.8)	12097(12.1)	12691 (2.3)	10723(10.7)
Diastolic blood pressure				
Mean (SD)	80.3 (10.4)	80.2 (11.1)	80.3 (10.4)	80.3 (11.1)
Missing	15386 (2.8)	12097(12.1)	12691 (2.3)	10723(10.7)
Atrial fibrillation				
No	545692(98.0)	97842(98.1)	544441(97.8)	97686(97.9)
Yes	10975 (2.0)	1941 (1.9)	12226 (2.2)	2097 (2.1)
Cancer				
No	529053(95.0)	91418(91.6)	525488(94.4)	90738(90.9)
Yes	27614 (5.0)	8365 (8.4)	31179 (5.6)	9045 (9.1)

All values are presented as n (%), unless otherwise stated.

	lookback period 7yr		lookback period 10yr	
Variables	Metformin	Sulfonylureas	Metformin	Sulfonylureas
Cerebrovascular disease				
No	528298(94.9)	92814(93.0)	525454(94.4)	92091(92.3)
Yes	28369 (5.1)	6969 (7.0)	31213 (5.6)	7692 (7.7)
Chronic Obstructive Pulmonary Disease				
No	530591(95.3)	95229(95.4)	529741(95.2)	94958(95.2)
Yes	26076 (4.7)	4554 (4.6)	26926 (4.8)	4825 (4.8)
Coronary heart disease				
No	473056(85.0)	81624(81.8)	466949(83.9)	80210(80.4)
Yes	83611 (15.0)	18159(18.2)	89718 (16.1)	19573(19.6)
Heart failure				
No	540960(97.2)	94039(94.2)	539657(96.9)	93606(93.8)
Yes	15707(2.8)	5744(5.8)	17010(1.7)	6177(6.2)
Hyperlipidemia				
No	462614(83.1)	87444(87.6)	448026(80.5)	85911(86.1)
Yes	94053 (16.9)	12339(12.4)	108641(19.5)	13872(13.9)
Hypertension				
No	307182(55.2)	64034(64.2)	294511(52.9)	61227(61.4)
Yes	249485(44.8)	35749(35.8)	262156(47.1)	38556(38.6)
Coronary Revascularization				
No	545209(97.9)	97642(97.9)	542415(97.4)	97244(97.5)
Yes	11458 (2.1)	2141 (2.1)	14252 (2.6)	2539 (2.5)
Thyroid				
No	520420(93.5)	94865(95.1)	515986(92.7)	94230(94.4)
Yes	36247 (6.5)	4918 (4.9)	40681 (7.3)	5553 (5.6)
Diabetic Neuropathy				
No	521721(93.7)	96524(96.7)	521143(93.6)	96463(96.7)
Yes	34946 (6.3)	3259 (3.3)	35524 (6.4)	3320 (3.3)
Peripheral Vascular Disease				
No	543243(97.6)	96381(96.6)	541824(97.3)	96008(96.2)
Yes	13424 (2.4)	3402 (3.4)	14843 (2.7)	3775 (3.8)
Chronic Kidney disease				
No	522375(93.8)	92786(93.0)	519647(93.3)	92496(92.7)
Yes	34292 (6.2)	6997 (7.0)	37020 (6.7)	7287 (7.3)
Retinal disorder				
No	528798(95.0)	96514(96.7)	528003(94.9)	96387(96.6)
Yes	27869 (5.0)	3269 (3.3)	28664 (5.1)	3396 (3.4)
eGFR				
Mean (SD)	81.7 (19.0)	69.4 (23.1)	81.8 (19.0)	69.5 (23.1)
Missing	60493 (10.9)	35429(35.5)	57736 (10.4)	35180(35.3)
Major Adverse Cardiovascular Events				
No	467178(83.9)	56843(57.0)	467178(83.9)	56843(57.0)
Yes	89489 (16.1)	4294(43.0)	89489 (16.1)	42940(43.0)
Previous stroke				
No	537557(96.6)	95490(95.7)	535581(96.2)	95033(95.2)
Yes	19110 (3.4)	4293 (4.3)	21086 (3.8)	4750 (4.8)
Previous Myocardial Infarction				
No	538430(96.7)	94785(95.0)	534400(96.0)	93895(94.1)
Yes	18237 (3.3)	4998 (5.0)	22267 (4.0)	5888 (5.9)

All values are presented as n (%), unless otherwise stated.

	lookback period 7yr		lookback period 10yr	
Variables	Metformin	Sulfonylureas	Metformin	Sulfonylureas
Medication				
Angiotensin Converting Enzyme (ACE)				
No	342371(61.5)	67889(68.0)	334374(60.1)	67093(67.2)
Yes	214296(38.5)	31894(32.0)	222293(39.9)	32690(32.8)
Angiotensin Receptor Blockers (ARB)				
No	487476(87.6)	91885(92.1)	485716(87.3)	91722(91.9)
Yes	69191 (12.4)	7898 (7.9)	70951 (12.7)	8061 (8.1)
Beta Blocker				
No	394152(70.8)	68731(68.9)	382933(68.8)	67226(67.4)
Yes	162515(29.2)	31052(31.1)	173734(31.2)	32557(32.6)
Diuretics				
No	365895(65.7)	60083(60.2)	356734(64.1)	59130(59.3)
Yes	190772(34.3)	39700(39.8)	199933(35.9)	40653(40.7)
Digoxin				
No	540215(97.0)	92023(92.2)	539757(97.0)	91934(92.1)
Yes	16452 (3.0)	7760 (7.8)	16910 (3.0)	7849 (7.9)
Statin				
No	253053(45.5)	68707(68.9)	251092(45.1)	68548(68.7)
Yes	303614(54.5)	31076(31.1)	305575(54.9)	31235(31.3)
Fibrate				
No	544644(97.8)	97529(97.7)	543242(97.6)	97336(97.5)
Yes	12023 (2.2)	2254 (2.3)	13425 (2.4)	2447 (2.5)
Aspirin				
No	392830(70.6)	68353(68.5)	386274(69.4)	67676(67.8)
Yes	163837(29.4)	31430(31.5)	170393(30.6)	32107(32.2)
Clopidogrel				
No	528064(94.9)	96023(96.2)	525939(94.5)	95917(96.1)
Yes	28603 (5.1)	3760 (3.8)	30728 (5.5)	3866 (3.9)
Warfarin				
No	528820(95.0)	92624(92.8)	527411(94.7)	92396(92.6)
Yes	27847 (5.0)	7159 (7.2)	29256 (5.3)	7387 (7.4)
Nonsteroidal				
Anti inflammatory				
Drugs (NSAIDS)				
No	226884(40.8)	50673(50.8)	199211(35.8)	47324(47.4)
Yes	329783(59.2)	49110(49.2)	357456(64.2)	52459(52.6)
Opioids				
No	252317(45.3)	50670(50.8)	230951(41.5)	48356(48.5)
Yes	304350(54.7)	49113(49.2)	325716(58.5)	51427(51.5)
Paracetamol				
No	254194(45.7)	49230(49.3)	233766(42.0)	47069(47.2)
yes	302473(54.3)	50553(50.7)	322901(58.0)	52714(52.8)
fibrate				
No	544644 (97.8)	97529 (97.7)	543242 (97.6)	97336 (97.5)
yes	12023 (2.2)	2254 (2.3)	13425 (2.4)	2447 (2.5)
Aspirin				
No	392830 (70.6)	68353 (68.5)	386274 (69.4)	67676 (67.8)
yes	163837 (29.4)	31430 (31.5)	170393 (30.6)	32107 (32.2)
clopidogrel				
No	528064 (94.9)	96023 (96.2)	525939 (94.5)	95917 (96.1)
yes	28603 (5.1)	3760 (3.8)	30728 (5.5)	3866 (3.9)
Warfarin				
No	528820 (95.0)	92624 (92.8)	527411 (94.7)	92396 (92.6)
yes	27847 (5.0)	7159 (7.2)	29256 (5.3)	7387 (7.4)

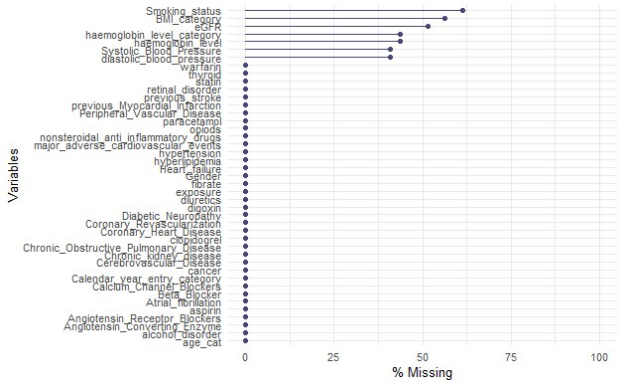
All values are presented as n (%), unless otherwise stated.

We presented the overall missing value percentage of variables in each lookback period (1,3, 6 months, and 5,7,10 years) in Figures 5.2a, 5.2b, 5.2c, 5.2d, 5.2e, 5.2f, 5.3a and 5.3b. The missing variables in each lookback are lifestyle variables (BMI and smoking status), hemoglobin level, eGFR, systolic blood pressure (sbp), and diastolic blood pressure(dbp).

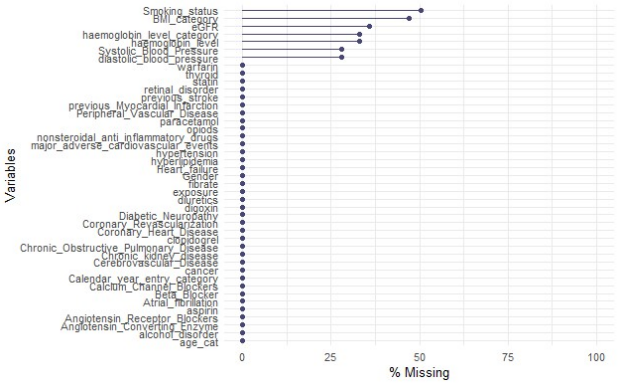
The overall percentage missing BMI for lookback periods 1 month, 3 months and 6 months is 62%, 51%, and 41% (Figure 5.2a, 5.2b, 5.2c). The missing percentage in the metformin group for the lookback period 1 month is 53.6% and 71.2% in the sulfonylureas group. This percentage decreases to 36.7% for metformin and 57% for the sulfonylureas group in lookback period of 6 months. The missing percentage of BMI for lookback period 1 year is 28% for metformin and 49.6% for the sulfonylureas group. The percentage decreases gradually as we increase the period of lookback length. For the lookback period of 5 years, the missing percentage for BMI is 12.1% for metformin and 31.8% (see Table 5.2) for the sulfonylureas group. As we increase the lookback period to 10 years, the percentage of missing BMI among metformin is 7.7%, and for sulfonylureas is 22.6% (see Table 5.2).

Similarly, for smoking status, the missing percentage for lookback 1 month is 62% (see Figure 5.2a) where the missing percentage for metformin group is 58.3% and for sulfonylureas 77.5% (see Table 5.1). By restricting lookback period to 1 year, these percentages decrease to 28% (25.4% for metformin and 51.3% for sulfonylureas), (see Table 5.2) and Figure 5.2d). The overall missing percentage for hemoglobin level for lookback period 1 month is 42% (62% (see Figure 5.2a) and decreases to 24% (Figure 5.2d), and it remains approximately the same for other lookback periods in years (5, 7, 9 yrs). The missing percentage for systolic and diastolic blood pressure for lookback 1 month is 40% and it decreases to 13% and 8% for lookback 1 and 2 years (see Figure 5.2d, 5.2e). This percentage decreases to 5% for lookback 5 years and less than 5% for lookback 7 and 10 years (see Figure 5.3a, 5.3b).

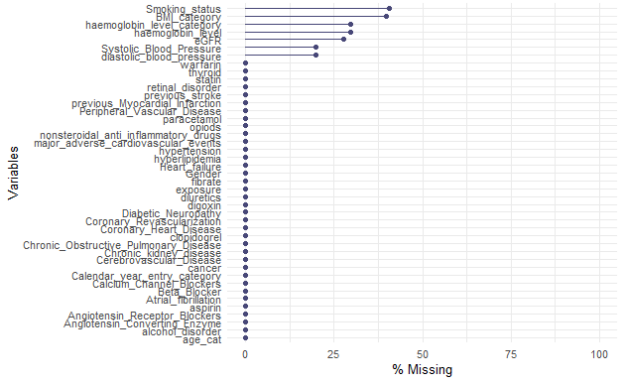
We analyzed data for each lookback period in two ways: complete case or available case analysis, and using multiple imputation. For simplicity, we kept our analysis plan the same for both approaches.



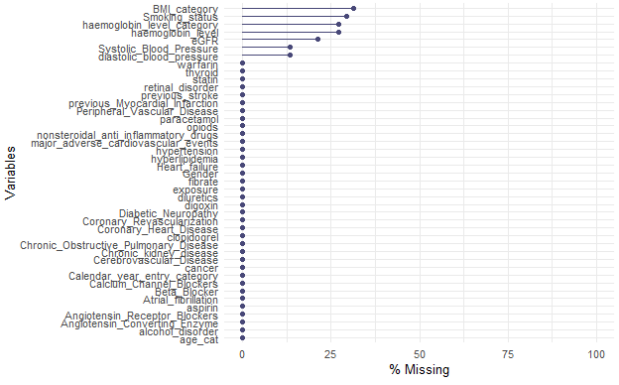
(a) Missing data plot for **lookback 1 month**



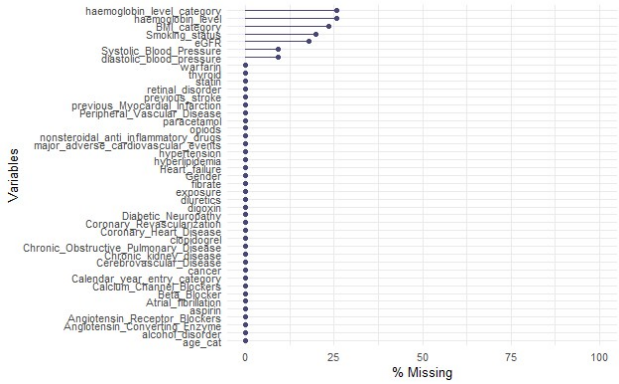
(b) Missing data plot for **lookback 3 months**



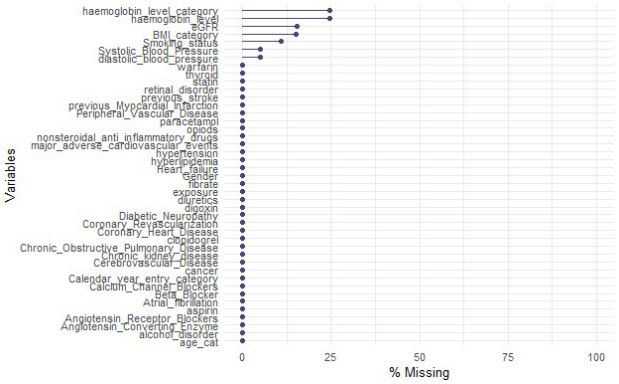
(c) Missing data plot for **lookback 6 months**



(d) Missing data plot for **lookback 1 year**

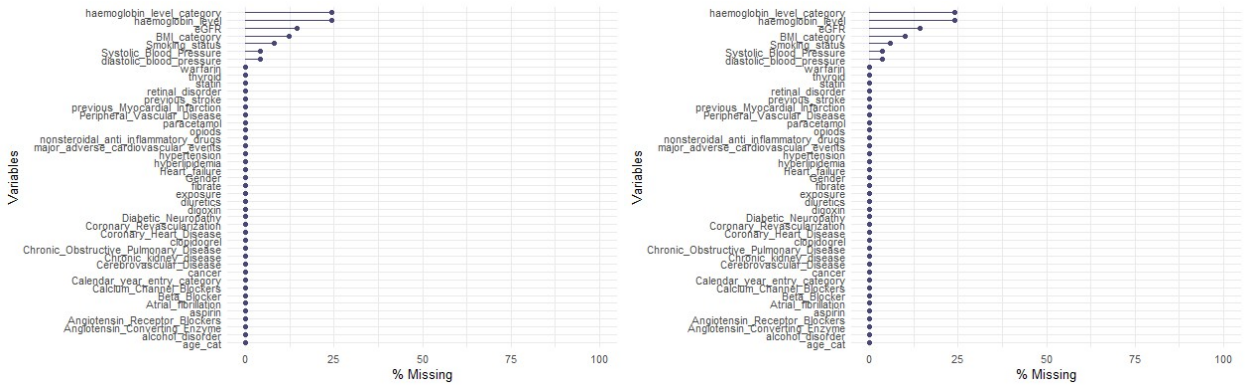


(e) Missing data plot for **lookback 2 year**



(f) Missing data plot for **lookback 5 year**

Figure 5.2: Missing data plots for varying lookback periods **(1, 3, 6 months and 1, 2, 5 years)**



(a) Missing data plot for **lookback 7 years** **(b)** Missing data plot for **lookback 10 years**

Figure 5.3: Missing data plots for varying lookback periods **(7 and 10 years)**

For the complete case analysis, we simply omit the participants with missing data (this could be for one variable only) and then analyzed the remaining data. For multiple imputations, we have used R package **mice** from the Comprehensive R Archive Network (CRAN)[141]. For the continuous missing variable sbp (systolic blood pressure), dbp (diastolic blood pressure), and eGFR, we used the **predictive mean matching (pmm)** method. Since smoking status is a binary categorical variable, we used **logreg (logistic regression)** [141]. For the ordered categorical missing variable BMI and hemoglobin level, we used **polr (proportional odds model or ordered logit model > two levels)** under the package **mice**. In each lookback, we imputed the data 5 times ($m = 5$ under mice package), and the total number of iterations was set to 20 (maxit=20 under mice). The above imputation methods are kept the same for all lookback periods.

We presented the **standardized mean difference (smd)** stratified by treatment group for varying lookback periods in Table 5.4. From the table, the smd seems to be > 0.1 for gender and age for each lookback. Among the comorbidities and risk factors, the variable alcohol disorder has smd smaller than 0.1 for lookback 1 month to 1 year and a higher smd (> 0.1) for lookback period of 2 to 10 years. The variables BMI, MACE, hemoglobin level, systolic blood pressure, cancer, hypertension, and eGFR have higher smd (> 0.1) in all lookback periods (Table 5.4). The variables heart failure, Diabetic Neuropathy have lower smd

in the lookback period of 1 and 3 months and higher smd in other lookback periods. The risk factor smoking status and other comorbidities including diastolic blood pressure, atrial fibrillation, cerebrovascular disease, chronic obstructive pulmonary disease, coronary heart disease, coronary revascularization, thyroid, peripheral vascular disease, chronic kidney disease, retinal disorder, history of stroke and myocardial infarction have lower smd (less than 0.1) in each lookback lengths. Only the covariate hyperlipidemia has a lower smd in lookback 1 month to 2 years and a higher smd for other lookbacks. The other covariates considered medication, i.e., ACE (angiotensin converting enzyme), ARB (angiotensin receptor blockers), diuretics, digoxin, and statin, have higher mean differences (smd > 0.1) for all lookback lengths. Only the medication NSAIDs has lower smd in lookback 1 month to 1 year, and it increases greater than 0.1 as the lookback period increases. The other medication variables, opioids and paracetamol have higher smd only at a lookback period of 10 years (see Table 5.4). So, covariates with important imbalances are present in the data.

Table 5.4: Comparison of **Standardized Mean Differences (SMD)** (Stratified by **Exposure**) Among Various Lookback Periods

	Standardized Mean Differences							
	Lookback Period							
	1mo	3mo	6mo	1yr	2yr	5yr	7yr	10yr
Variables								
Gender= Male	0.062	0.062	0.062	0.062	0.062	0.062	0.062	0.062
Age	0.411	0.411	0.411	0.411	0.411	0.411	0.411	0.411
Calendar year entry category	1.189	1.189	1.189	1.189	1.189	1.189	1.189	1.189
Comorbidities and risk actors								
BMI category	0.898	0.085	0.829	0.797	0.765	0.719	0.705	0.70
Major adverse cardiovascular events (MACE)	0.618	0.618	0.618	0.618	0.618	0.618	0.618	0.618
Smoking status	0.048	0.044	0.042	0.038	0.039	0.057	0.072	0.083
Alcohol disorder	0.070	0.071	0.073	0.083	0.109	0.146	0.161	0.175
Hemoglobin level category	0.518	0.483	0.434	0.381	0.355	0.34	0.337	0.334
Systolic blood pressure	0.134	0.145	0.158	0.173	0.187	0.201	0.201	0.201
Diastolic blood pressure	0.060	0.057	0.049	0.041	0.03	0.012	0.007	0.004
Atrial fibrillation	0.009	0.011	0.013	0.015	0.009	0.001	0.002	0.007
Cancer	0.126	0.148	0.158	0.156	0.153	0.141	0.137	0.133
Cerebrovascular disease	0.030	0.039	0.050	0.055	0.062	0.071	0.079	0.084
Chronic Obstructive Pulmonary Disease	0.011	0.014	0.015	0.021	0.019	0.01	0.006	0.001
Coronary heart disease	0.024	0.031	0.037	0.045	0.058	0.077	0.085	0.091
Heart failure	0.072	0.090	0.104	0.114	0.124	0.141	0.145	0.15
Hyperlipidemia	0.053	0.050	0.047	0.05	0.065	0.109	0.128	0.151
Hypertension	0.135	0.153	0.157	0.17	0.185	0.192	0.184	0.171
Coronary revascularization	0.038	0.039	0.034	0.028	0.021	0.011	0.006	0.001
Thyroid	0.037	0.014	0.040	0.047	0.05	0.064	0.068	0.071
Peripheral Vascular Disease	0.016	0.023	0.031	0.039	0.045	0.054	0.059	0.063
Diabetic Neuropathy	0.092	0.094	0.102	0.115	0.122	0.138	0.142	0.142
Chronic kidney disease	0.016	0.023	0.031	0.055	0.055	0.042	0.034	0.026
Retinal disorder	0.017	0.036	0.046	0.066	0.075	0.085	0.087	0.086
eGFR	0.528	0.549	0.558	0.568	0.575	0.58	0.582	0.582
Previous stroke	0.026	0.029	0.035	0.035	0.037	0.04	0.045	0.048
Previous Myocardial Infarction	0.069	0.077	0.079	0.08	0.077	0.084	0.087	0.088
Medications								
Angiotensin converting enzyme	0.115	0.135	0.127	0.12	0.117	0.13	0.137	0.15
Angiotensin receptor blockers	0.139	0.158	0.152	0.147	0.143	0.147	0.15	0.153
Beta blocker	0.024	0.030	0.041	0.048	0.052	0.049	0.042	0.03
Diuretics	0.113	0.124	0.136	0.14	0.139	0.124	0.114	0.099
Digoxin	0.182	0.204	0.209	0.211	0.213	0.215	0.215	0.214
Statin	0.405	0.466	0.468	0.469	0.472	0.482	0.487	0.49
Fibrate	0.005	0.006	0.001	0.003	0.007	0.01	0.007	0.003
aspirin	0.043	0.053	0.066	0.069	0.069	0.056	0.045	0.034
Clopidogrel	0.015	0.023	0.025	0.027	0.035	0.057	0.066	0.078
Warfarin	0.067	0.083	0.091	0.095	0.096	0.092	0.091	0.088
Nonsteroidal Anti-Inflammatory Drugs	0.031	0.038	0.045	0.07	0.105	0.174	0.202	0.238
Opioids	0.025	0.030	0.029	0.014	0.015	0.079	0.109	0.141
Paracetamol	0.025	0.030	0.029	0.048	0.02	0.044	0.074	0.104

We kept our analysis approach similar to our simulation studies. We estimated the overall effect using the propensity score method under the IPTW framework and the TMLE approach. For Propensity score methods under inverse probability treatment weighting framework, we computed the distribution of propensity score distribution for both complete case and imputed analysis. Table 5.5 and Table 5.6 present the summary statistics for propensity score distribution under varying lookback lengths. From the complete-case analysis, the median of PS estimates in the metformin group ranges between 0.028 to 0.037, and for the sulfonylureas group, it ranges from 0.169 to 0.203. The mean of propensity score estimates for varying lookback length ranges between 0.059 to 0.074 for the metformin and 0.235 to 0.266 for the sulfonylureas group. However, the estimates of 3rd quartiles differ between the two groups (0.064 – 0.081) for metformin and (0.361 – 0.403) for sulfonylureas but do not vary much among the lookback periods in each group. From the imputed analysis, the mean of PS estimates for the metformin group ranges from 0.117 to 0.112 and 0.347 – 0.370 from the sulfonylureas group. Also, for imputed analysis, the quantile estimates, minimum and maximum, were not affected by varying lookback lengths (see Table 5.6).

We have used the **stabilized weight** for both analyses (complete and imputed analysis) to reduce influential weights. We presented the distribution of the stabilized weight in Table 5.7 and Table 5.8 for varying lookback periods.

Table 5.5: Propensity Score Distribution Under Varying Lookback Periods for Complete Case Analysis

	Metformin						**Sulfonylureas**					
Lookback	min	1st q	Median	Mean	3rd q	Max	min	1st q	Median	Mean	3rd q	Max
1mo	0.0014	0.013	0.028	0.059	0.064	0.912	0.0034	0.070	0.169	0.235	0.343	0.94
3mo	0.0011	0.013	0.027	0.059	0.063	0.933	0.002	0.071	0.174	0.239	0.350	0.96
6mo	0.0014	0.015	0.033	0.062	0.068	0.941	0.003	0.075	0.178	0.243	0.354	0.95
1yr	0	0.016	0.032	0.066	0.073	0.95	0.003	0.077	0.182	0.246	0.361	0.95
2yr	0	0.017	0.034	0.068	0.075	0.95	0.003	0.079	0.189	0.252	0.376	0.95
5yr	0	0.019	0.036	0.071	0.078	0.95	0.004	0.081	0.195	0.25	0.389	0.96
7yr	0	0.019	0.036	0.072	0.079	0.94	0.005	0.083	0.19	0.26	0.398	0.96
10yr	0	0.02	0.037	0.074	0.081	0.94	0.005	0.084	0.203	0.266	0.403	0.95

After weighting, we presented **smd (standardized mean difference) plots** for all lookback periods for both cases under the PS-IPW framework to check the covariate bal-

Table 5.6: Propensity Score Distribution Under Varying Lookback Periods for Imputed Analysis

	Metformin						Sulfonylureas					
Lookback	Min	1st q	Median	Mean	3rd q	Max	min	1st q	Median	Mean	3rd q	Max
1mo	0.004	0.036	0.061	0.117	0.122	0.959	0.007	0.134	0.358	0.347	0.516	0.972
3mo	0	0.034	0.059	0.115	0.121	0.961	0.006	0.143	0.358	0.354	0.526	0.982
6mo	0	0.033	0.058	0.115	0.121	0.963	0.007	0.146	0.358	0.357	0.530	0.990
1yr	0	0.032	0.057	0.114	0.121	0.962	0.006	0.149	0.36	0.360	0.533	0.986
2yr	0	0.0314	0.056	0.114	0.122	0.962	0.006	0.153	0.361	0.363	0.536	0.971
5yr	0	0.030	0.055	0.113	0.123	0.961	0.005	0.156	0.361	0.367	0.544	0.97
7yr	0	0.0298	0.055	0.113	0.123	0.949	0.006	0.156	0.361	0.369	0.547	0.971
10yr	0	0.029	0.055	0.112	0.124	0.952	0.005	0.157	0.362	0.370	0.552	0.96

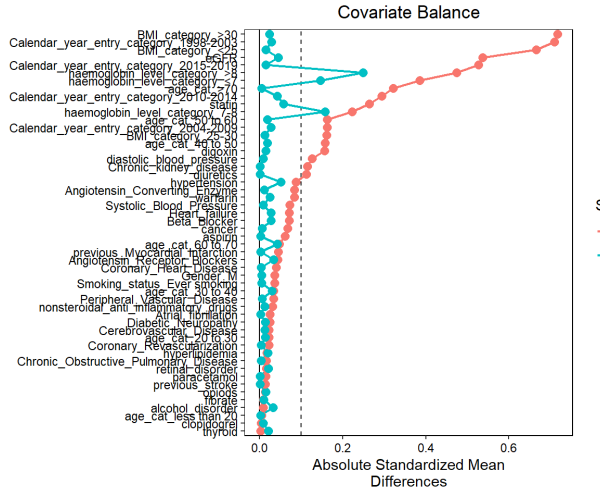
Table 5.7: Distribution of Stabilized Weight Under IPW for Varying Lookback Periods Under Complete Case Analysis

Lookback	Min	1st q	Median	Mean	3rd q	Max
1mo	0.07567	0.93985	0.95398	0.99601	0.99192	20.7137
3mo	0.07528	0.93921	0.95298	0.99731	0.99091	28.8051
6mo	0.08006	0.93603	0.95076	0.99775	0.99165	24.4416
1yr	0.08444	0.93312	0.94884	0.99852	0.99222	25.0361
2yr	0.08799	0.93044	0.94649	0.99852	0.99117	22.6512
5yr	0.09148	0.92796	0.94437	0.99862	0.99032	18.9099
7yr	0.09324	0.92634	0.94286	0.99852	0.98948	17.8236
10yr	0.09572	0.92476	0.94161	0.99847	0.98932	17.5944

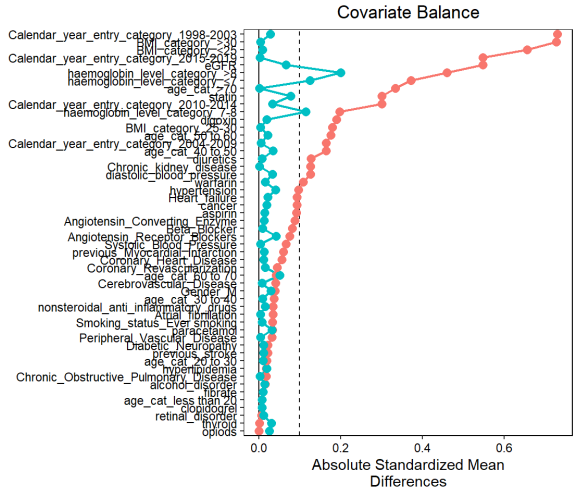
Table 5.8: Distribution of Stabilized Weight Under IPW for Varying Lookback Periods With Imputed Analysis

Lookback	Min	1st q	Median	Mean	3rd q	Max
1mo	0.1563	0.8740	0.8993	0.9971	0.9686	20.8448
3mo	0.1547	0.8721	0.8969	0.9966	0.9674	24.4718
6mo	0.1535	0.8714	0.8957	0.9964	0.9668	23.3269
1yr	0.1542	0.8705	0.8946	0.9965	0.9666	24.8408
2yr	0.1564	0.8698	0.8937	0.9968	0.9668	23.4751
5yr	0.1565	0.8689	0.8927	0.9973	0.9675	25.6681
7yr	0.1565	0.8685	0.8924	0.9978	0.9680	22.3650
10yr	0.1571	0.8681	0.8921	0.9978	0.9682	27.4739

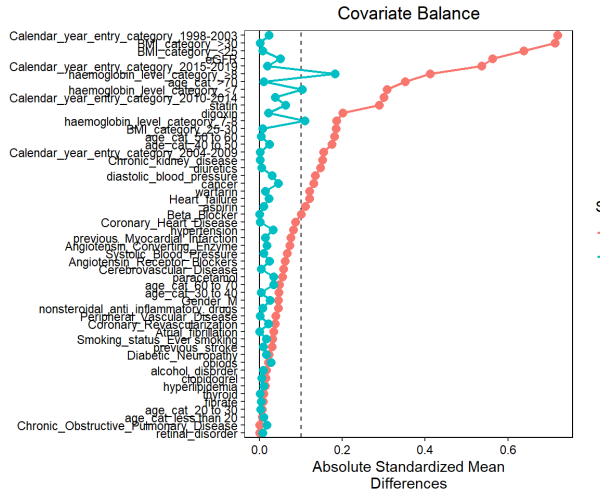
ance. Figure 5.4a to 5.5d represents the covariate balance before and after weighting under complete case scenarios for varying lookback lengths. The smd plots for imputed analysis with varying lookback lengths are presented from Figure 5.6a to 5.7d.



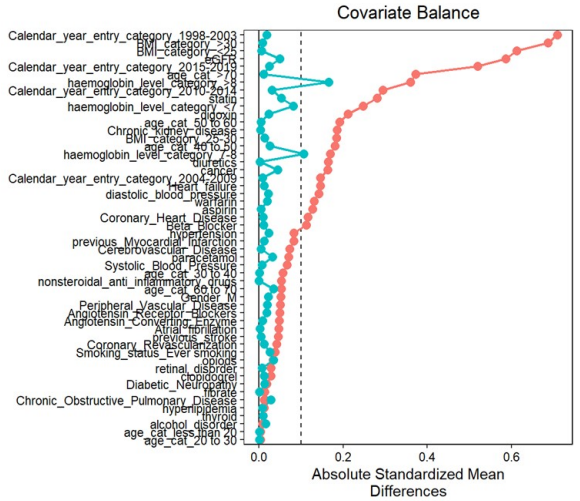
(a) Comparison of standardized mean difference before and after weighting for **lookback 1 month**



(b) Comparison of standardized mean difference before and after weighting for **lookback 3 months**

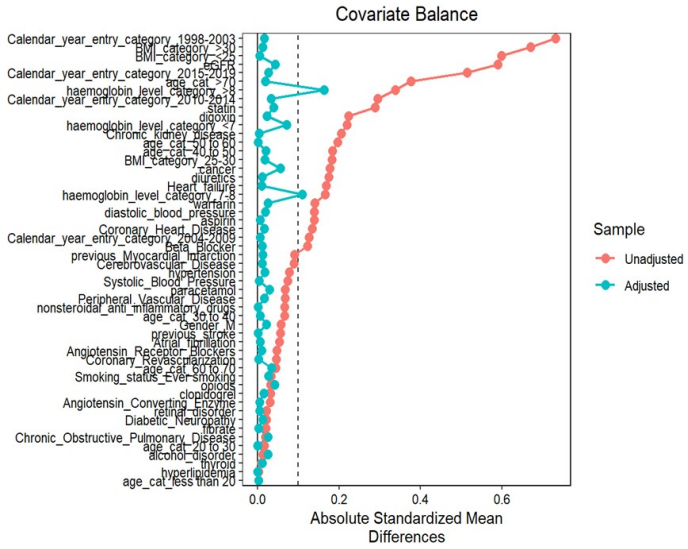


(c) Comparison of standardized mean difference before and after weighting for **lookback 6 months**

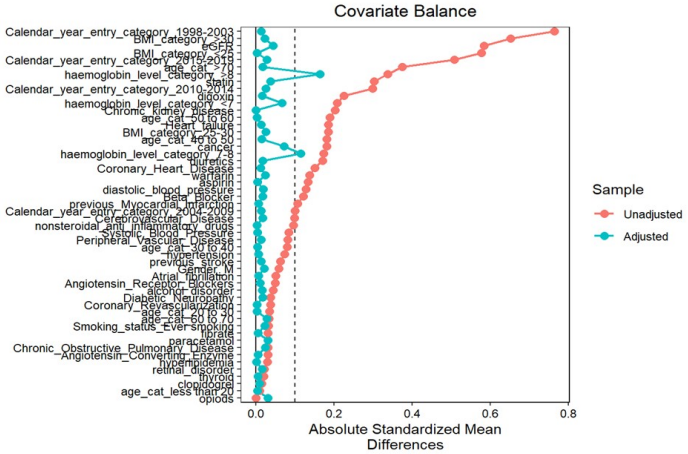


(d) Comparison of standardized mean difference before and after weighting for **lookback 1 year**

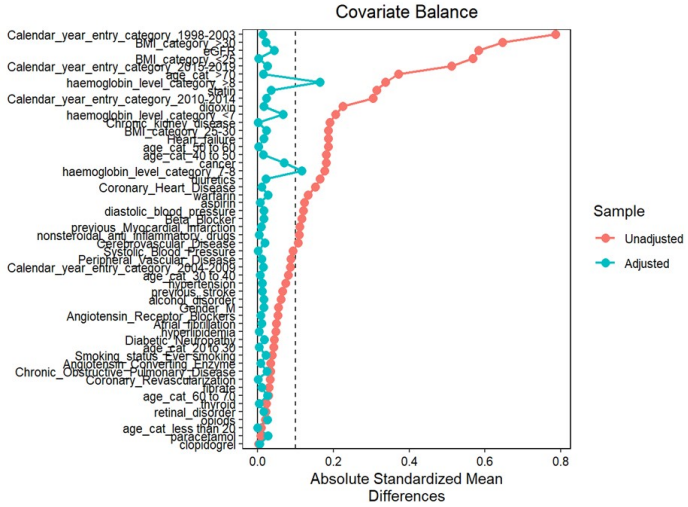
Figure 5.4: Comparison of standardized mean difference before and after weighting for varying lookback periods **(1, 3, 6 months and 1 year)**



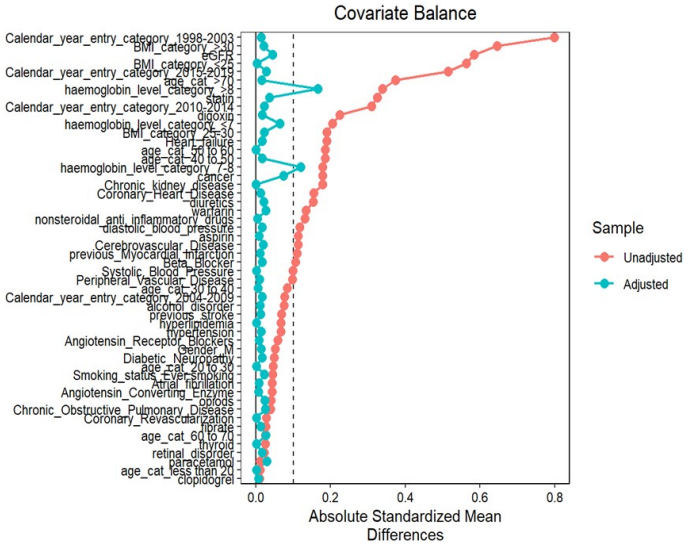
(a) Comparison of standardized mean difference before and after weighting for **lookback 2 years**



(b) Comparison of standardized mean difference before and after weighting for **lookback 5 years**

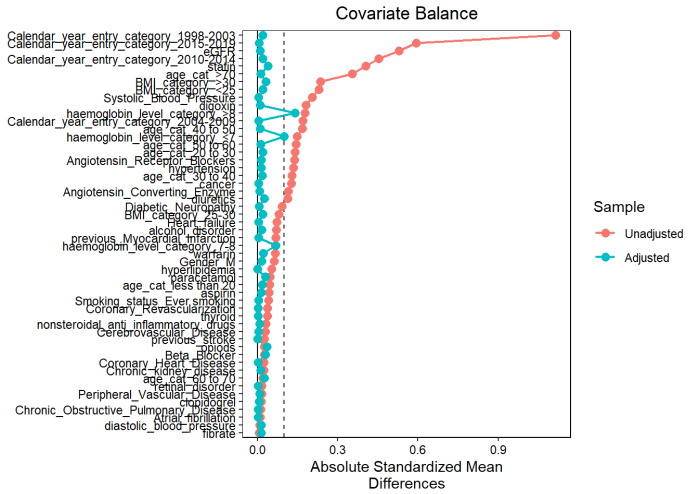


(c) Comparison of standardized mean difference before and after weighting for **lookback 7 years**

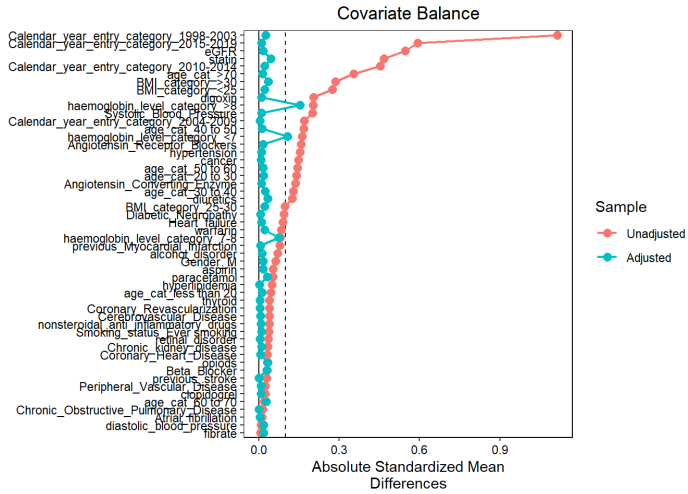


(d) Comparison of standardized mean difference before and after weighting for **lookback 10 years**

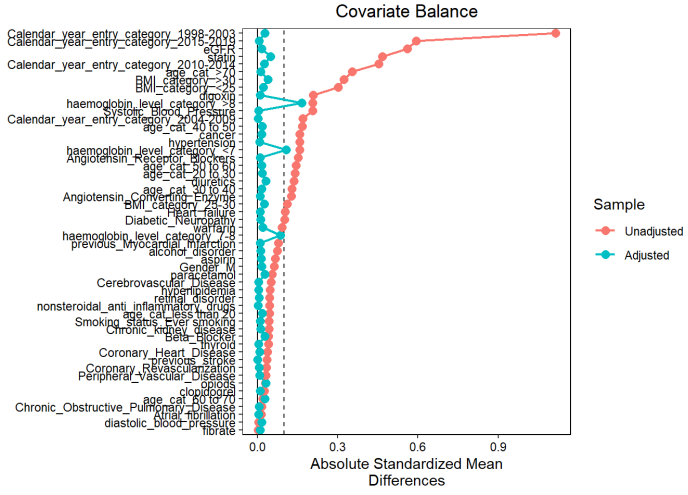
Figure 5.5: Comparison of standardized mean difference before and after weighting for varying lookback periods **(2, 5, 7, and 10 years)**



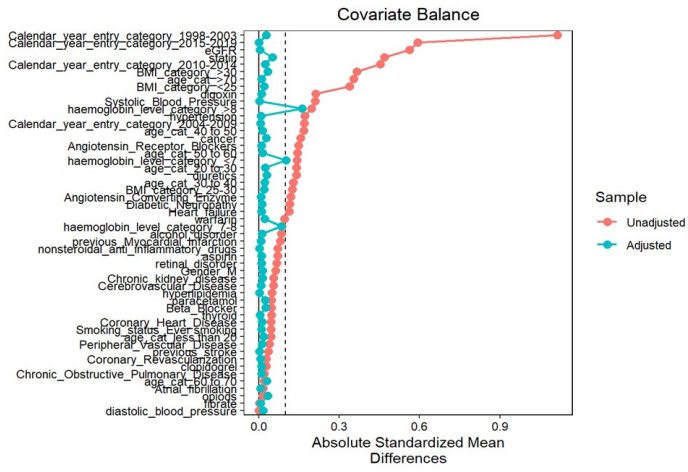
(a) Comparison of standardized mean difference before and after weighting for **lookback 1 month** for imputed analysis



(b) Comparison of standardized mean difference before and after weighting for **lookback 3 months** for imputed analysis

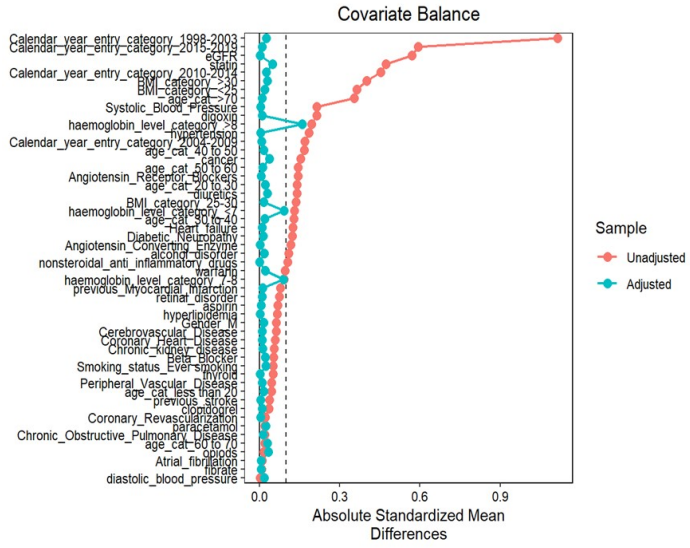


(c) Comparison of standardized mean difference before and after weighting for **lookback 6 months** for imputed analysis

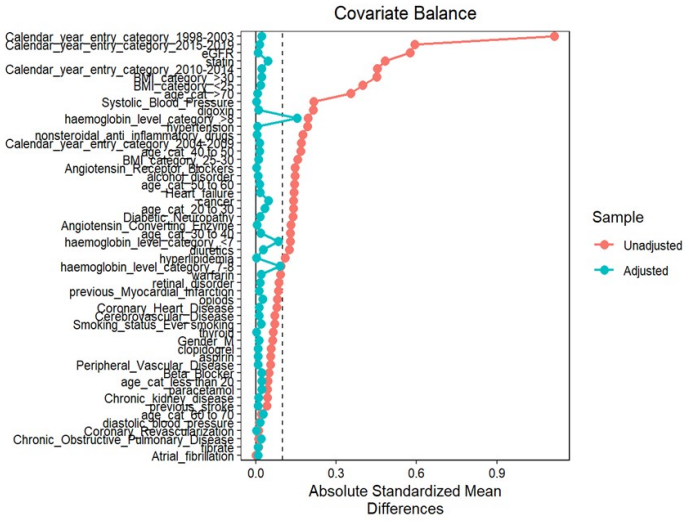


(d) Comparison of standardized mean difference before and after weighting for **lookback 1 year** for imputed analysis

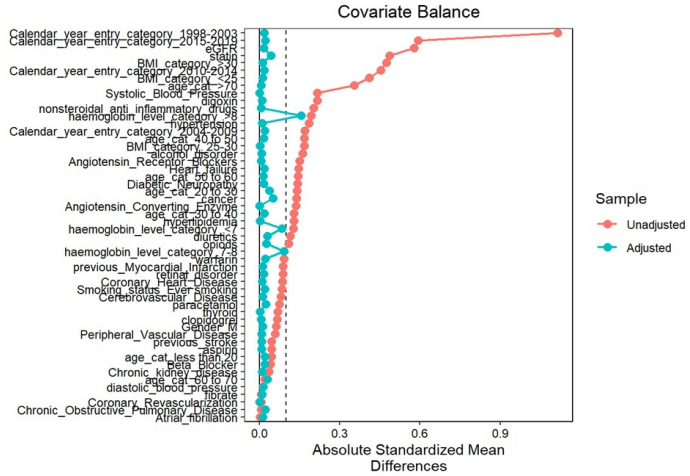
Figure 5.6: Comparison of standardized mean difference before and after weighting for varying lookback periods **1, 3, 6 months and 1 year** for imputed analysis



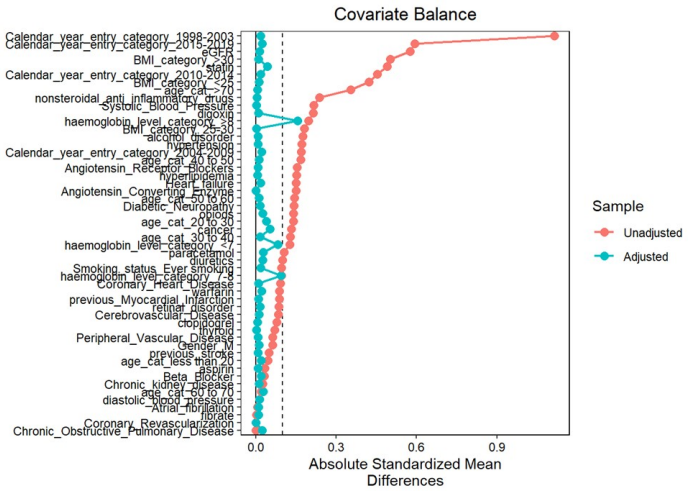
(a) Comparison of standardized mean difference before and after weighting for **lookback 2 years** for imputed analysis



(b) Comparison of standardized mean difference before and after weighting for **lookback 5 years** for imputed analysis



(c) Comparison of standardized mean difference before and after weighting for **lookback 7 years** for imputed analysis



(d) Comparison of standardized mean difference before and after weighting for **lookback 10 years** for imputed analysis

Figure 5.7: Comparison of standardized mean difference before and after weighting for varying lookback periods **2, 5, 7 and 10 years** for imputed analysis

5.2 Summary

Following the methods of the primary study, our analyses use an intention-to-treat approach to compare our outcomes among the patients who initiated sulfonylureas and among the patients who initiated metformin. Using the targeted maximum likelihood estimation approach and the propensity score method in inverse probability weighting framework, the effect of treatment is estimated by restricting lookback periods correspondingly to lookback 1, 3, 6 months, and 7, 5, 2, 1 years.

We present the effect estimates of sulfonylureas use on major adverse cardiovascular events (MACE) for various lookback lengths by TMLE and PS-IPW approach in Table 5.9 and Table 5.10 for complete and imputed analysis. The standard errors should account for the weighting, so as in our simulation studies, we used the robust standard errors for ATE using the IPTW method [82]. To do this in R, we used survey::vcovHC() after a glm() call with the outcome model recommended by [82]. In both exposure and outcome models, we considered all the covariates considered in Tables 5.1, 5.2, 5.3 for each lookback analysis.

For TMLE, we used the same propensity score model as for the PS-IPW approach. We considered the outcome model including all adjusted confounders and estimated the ATE. We presented ATE for all lookbacks and corresponding standard errors in Table 5.9 for the complete case and Table 5.10 for imputed analysis. We also estimated the log odds ratio under this approach for different lookbacks. For both PS-IPW and TMLE-logistic, no higher-order polynomials or interaction term was added. The standard errors for the estimates for the TMLE were calculated based on the idea of efficient influence curve (EIC) [25, 58]. We also estimated the log odds ratio and corresponding standard error using TMLE and IPW with varying lookback periods for both imputed and complete case analysis.

Table 5.9: Effect Estimates of Sulfonylureas Using Logistic Regression Model (Adjusted for all Covariates) for TMLE and PS (IPW-Stabilized Weights) Under Various Lookbacks **(Complete Case)**

Lookback Period		ATE	SE(ATE)	log (OR)	SE (log (OR))
10yr		0.0718	0.0028	0.6217	0.0149
7yr		0.0713	0.0029	0.6177	0.0153
5yr	PS-IPW(Stabilized)	0.0696	0.0029	0.6028	0.0157
2yr		0.0623	0.0031	0.5355	0.0173
1yr		0.0582	0.0036	0.4903	0.0193
6mo		0.0587	0.0043	0.4834	0.0228
3mo		0.0555	0.0049	0.4509	0.0274
1mo		0.0516	0.0060	0.3928	0.0349
10yr		0.0724	0.0026	0.4660	0.0153
7yr		0.0718	0.0027	0.4635	0.0158
5yr	TMLE-logistic	0.0701	0.0028	0.4530	0.0162
2yr		0.0628	0.0030	0.4059	0.0177
1yr		0.0577	0.0034	0.3746	0.0203
6mo		0.0571	0.0040	0.3759	0.0241
3mo		0.0524	0.0047	0.3571	0.2912
1mo		0.0467	0.0055	0.3224	0.0348

Table 5.10: Effect Estimates of Sulfonylureas Using Logistic Regression Model (Adjusted for all Covariates) for TMLE and PS (IPW-Stabilized weights) Under Various Lookbacks **(With Imputation)**

Lookback Period		ATE	SE(ATE)	log (OR)	SE (log (OR))
10yr		0.0722	0.00180	0.5766	0.00960
7yr		0.0713	0.00185	0.5710	0.00961
5yr	PS-IPW(Stabilized)	0.0704	0.00183	0.5634	0.00963
2yr		0.0678	0.00182	0.5433	0.00965
1yr		0.0685	0.00180	0.5439	0.00964
6mo		0.0685	0.00178	0.5415	0.00963
3mo		0.0721	0.00179	0.5647	0.00958
1mo		0.0821	0.00181	0.6257	0.00946
10yr		0.0738	0.00178	0.4252	0.00943
7yr		0.0733	0.00176	0.4209	0.00935
5yr	TMLE-logistic	0.0722	0.00173	0.4166	0.00921
2yr		0.0695	0.00170	0.4023	0.00911
1yr		0.0698	0.00169	0.4041	0.00906
6mo		0.0696	0.00166	0.4021	0.00881
3mo		0.0729	0.00168	0.4205	0.00892
1mo		0.0822	0.00170	0.4702	0.00900

Chapter 6

Discussion

The use of population health data to study the nature of health outcomes research is growing [142, 143]. However, the incomplete nature of the information in administrative databases about patients' comorbidities and history raises an important concern regarding the lookback period. Few studies have shown the impact of longer lookback length compared to shorter lookback length, and almost all of them focused on using propensity score methods to adjust for confounders. However, to our knowledge, the doubly robust method TMLE has not been implemented in the administrative databases with a variety of lookback periods. In this study, we implemented the doubly robust method TMLE on different lookback data and made a comparison between propensity score and TMLE approach using simulation and a real life database.

6.1 Summary Findings from Simulation Study

Our simulation result shows an important variation in the quality of the PS estimates for short term lookback (1, 3, 6 months and 1, 2 years) compared to longer lookback (5, 7 and 9 years) as expected. We considered 10 years lookback as an ideal scenario to compare the PS quantile estimates for various lookback periods. The longer lookback length has less biased estimates compared to the shorter lookbacks. And the estimates of propensity score distribution for a lookback period of 9 years produce very close results to the ideal scenario. These findings underscore the importance of using longer lookback time windows

for more reliable estimates, and we recommend their adoption in applications after carefully addressing issues related to missingness and loss to follow-up. Relevant variations were found among the estimates of ATE and standard errors for varying lookback periods. The average effect estimates for longer lookback periods (9, 7 and 5 years) were close to the true value than for every three methods (TMLE-logistic, PS-IPW, TMLE-SL) compared to the shorter lookback periods. Furthermore, the doubly robust method TMLE, exhibited greater efficiency in producing smaller standard errors than propensity score methods, particularly for longer lookback periods (9, 7, and 5 years). Incorporating SuperLearner with TMLE proved effective, resulting in smaller standard errors for ATE estimates when compared to TMLE-logistic and PS-IPW. These results highlight the advantages of the data-adaptive approach through TMLE, offering valid inference and providing a compelling incentive for applied practitioners to utilize TMLE in their research.

6.2 Summary Findings from the Analysis of Real Data

In our real world data analysis, we evaluated the covariates with a lookback of 10 years and then restricted the period to 1, 3, 6 months and 1, 2, 5, 7 years. Missing percentages for comorbidities were higher when restricting the lookback period to 1 month to 1 year. But these percentages did not vary much for lookback 1 year to 10 years. This could happen because, for shorter lookback (i.e., 1 month to 6 months), much information on the patient's comorbidities was not available in the database. So we might not expect much variation in the effect estimates among the various lookback 1 to 10 years. But if we look at lookback 1 month to 1 year, we can see some differences among the estimates of treatment effect and standard errors.

In the complete case analysis, the effect estimates did not differ much among the varying lookback lengths for PS-IPW (stabilized) and logistic TMLE. But we can see the estimates of the standard errors for ATE varied from lookback 1 month to 1 year. In further lookbacks (1 year to 10 years), we could see less variation among the estimates of standard errors for both TMLE and IPW methods. For the imputed analysis, the average effect estimates

did not vary much among the lookbacks for the two methods. The estimates of standard errors among different lookbacks (1 month to 1 year) were very close to each other (both for IPW and TMLE). This is expected because, in imputed analysis, no incompleteness of information is present. So, it is very likely to have similar results for each lookback period. However, the standard error estimates of average treatment effects for imputed analysis are less compared to the complete case analysis, as expected. In both the complete case and imputed analyses, the logistic TMLE has smaller standard errors for ATE estimates than stabilized PS-IPW for each lookback length. For the imputed analysis, the estimated standard errors of the log odds ratio were smaller for TMLE-logistic, whereas, for the complete case analysis, PS-IPW performed well.

Both in real data and simulation study, TMLE-logistic did not produce a great reduction in variance, but this new doubly robust method performs comparatively well for each lookback for estimating the average treatment effect. In our real analysis, we were unable to assess the bias of the two methods as the true effect was unknown.

6.3 Strengths of the Study

Our study delved into the variability of average treatment effect estimates and standard errors across a wide range of lookback periods (ranging from 1 month to 10 years) in observational study settings. We evaluated these various lookback periods using PS-IPW, TMLE-logistic, and TMLE-SL, comparing the effect estimates through both simulation and real data analyses. In real data settings, the estimates of PS-quantiles remained unaffected by the varying lookback periods, mirroring real world scenarios. Conversely, in controlled simulation conditions, the quantiles of PS estimates exhibited variations among the different lookback periods. In both the simulation and real data analyses, we found that the double robust TMLE outperformed the PS method (PS-IPW), with longer lookback periods leading to lower standard error estimates. Notably, our simulation study favored integrating TMLE with an ensemble machine learning algorithm for varying lookback scenarios, highlighting its superiority over the parametric TMLE and PS-IPW approach. Furthermore, while pre-

senting the methodological challenges, the extended lookback periods allowed us to capture the full history of the patient's health in detail and made our findings more reliable. Despite having complexities, this strategic choice facilitated a broad evaluation of the temporal impact of comorbidities on health outcomes.

6.4 Limitations of the Study

After restricting the lookback periods in our analysis, all covariates remained fixed, which means these results cannot be generalized to scenarios with time-varying covariates. Furthermore, it's important to note that our simulation study may not fully capture all possible lookback settings. Therefore, we cannot make definitive conclusions regarding the comparison of TMLE and PS-IPW in other lookback scenarios. Moreover, in our real analysis, we demonstrated the implementation of TMLE in a simpler scenario. However, we encountered computational constraints that prevented us from integrating TMLE with SuperLearner into the real data analysis. In the real data analysis, we observed that the proportion of missing values was higher while restricting the lookback to shorter periods. This finding is substantial as it highlights the potential for biased estimates of different treatment effects due to incomplete data. Despite this, we did not see much variation in the effect estimates of varying lookbacks (1-10 years). This suggests that the imputation method we applied may mitigate the impact of missing data for the more extended lookback periods. Nonetheless, we need to be careful in interpreting these findings, as the assumption of the imputation model may not hold in all scenarios.

6.5 Future Study

Future studies should be done on other lookback settings and consider the different model scenarios, such as non-additivity, non-linearity, and other complex model scenarios. Collaborative TMLE (C-TMLE), an extension of TMLE [144, 145], may be useful in the settings we studied. Future work should consider these and other alternative TMLE methods. Additionally, more work should be on implementing TMLE with SuperLearner with a large number

of covariates for restricted lookback periods in real-world scenarios. This may outperform the parametric TMLE in the presence of more complex data. Future research should also aim to incorporate more sophisticated methods for dealing with missing data and perform sensitivity analysis to provide a more nuanced understanding of the impact of missingness on study outcomes.

6.6 Conclusion

In this thesis, we address the criticism surrounding administrative healthcare data for potential confounding by determining an appropriate lookback period for assessing confounders relative to exposure. We demonstrated the implementation of double robust TMLE and propensity score methods using simulation for varying lookback periods in observational settings as well as in real data to estimate the treatment effect of sulfonylureas on major adverse cardiovascular events. To achieve this, we conduct simulation studies in observational settings and analyze real data. Both TMLE and propensity score under inverse probability weighting framework (PS-IPW) are implemented using the same modeling approach, allowing us to compare their effect estimates and standard errors. In the simulation, we observe that TMLE-logistic yields lower standard errors for effect estimates compared to PS-IPTW in scenarios with longer lookback periods. Conversely, in scenarios with shorter lookback periods, the variation in standard error between TMLE and PS-IPW is less pronounced. Additionally, when we apply TMLE with SuperLearner to varying lookback scenarios, we find that it outperforms both TMLE with logistic regression and PS-IPW. However, when analyzing real data, we observed smaller standard errors for effect estimates for all long and short term lookback periods. This study will aid researchers in understanding the implications of the TMLE method when applied to diverse lookback periods in observational data, thereby ensuring the validity and generalizability of forthcoming study results.

Bibliography

[1] John Kendall. “Designing a research project: randomised controlled trials and their principles”. In: *Emergency medicine journal: EMJ* 20.2 (2003), p. 164.

[2] Cecilia Nardini. “The ethics of clinical trials”. In: *Ecancermedicalscience* 8 (2014).

[3] Jennifer A Hutcheon et al. “Good practices for the design, analysis, and interpretation of observational studies on birth spacing and perinatal health outcomes”. In: *Paediatric and Perinatal Epidemiology* 33.1 (2019), O15–O24.

[4] David B Resnik. “The precautionary principle and medical decision making”. In: *The Journal of medicine and philosophy* 29.3 (2004), pp. 281–299.

[5] Raanan Gillon. “Medical ethics: four principles plus attention to scope”. In: *Bmj* 309.6948 (1994), p. 184.

[6] Anat Fisher et al. “Patterns of antiemetic medication use during pregnancy: A multi-country retrospective cohort study”. In: *Plos one* 17.12 (2022), e0277623.

[7] Robert W Platt, D Henry, and Samy Suissa. “The Canadian Network for Observational Drug Effect Studies (CNODES): reflections on the first eight years, and a look to the future”. In: (2019).

[8] Samy Suissa et al. “CNODES: the Canadian network for observational drug effect studies”. In: *Open Medicine* 6.4 (2012), e134.

[9] Robert W Platt et al. “Observational studies of drug safety in multi-database studies: methodological challenges and opportunities”. In: *eGEMs* 4.1 (2016).

[10] Robert W Platt et al. *How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias*. 2020.

[11] Kristian B Filion et al. “A multicenter observational study of incretin-based drugs and heart failure”. In: *New England Journal of Medicine* 374.12 (2016), pp. 1145–1154.

[12] Christine Y Lu. “Observational studies: a review of study designs, challenges and strategies to reduce confounding”. In: *International journal of clinical practice* 63.5 (2009), pp. 691–697.

[13] Donald B Rubin. “Propensity score methods”. In: *American journal of ophthalmology* 149.1 (2010), pp. 7–9.

[14] Ralph B D’Agostino Jr. “Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group”. In: *Statistics in medicine* 17.19 (1998), pp. 2265–2281.

[15] Haiyan Bai and Margaret Hilary Clark. *Propensity score methods and applications*. Sage Publications, 2018.

[16] Peter C Austin. “An introduction to propensity score methods for reducing the effects of confounding in observational studies”. In: *Multivariate behavioral research* 46.3 (2011), pp. 399–424.

[17] Susan M Perkins et al. “The use of propensity scores in pharmacoepidemiologic research”. In: *Pharmacoepidemiology and drug safety* 9.2 (2000), pp. 93–101.

[18] Paul R Rosenbaum. “Model-based direct adjustment”. In: *Journal of the American Statistical Association* 82.398 (1987), pp. 387–394.

[19] Steven M Brunelli et al. “Estimation using all available covariate information versus a fixed look-back window for dichotomous covariates”. In: *Pharmacoepidemiology and drug safety* 22.5 (2013), pp. 542–550.

[20] Sonja S Nakasian, Jeremy A Rassen, and Jessica M Franklin. “Effects of expanding the look-back period to all available data in the assessment of covariates”. In: *Pharmacoepidemiology and Drug Safety* 26.8 (2017), pp. 890–899.

[21] FDA FDA Draft Guidance. “Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data Sets”. In: *FDA Draft Guidance* (2011).

[22] Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*. Vol. 10. Springer, 2011.

[23] Jared K Lunceford and Marie Davidian. “Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study”. In: *Statistics in medicine* 23.19 (2004), pp. 2937–2960.

[24] Peter Bühlmann et al. *Handbook of big data*. CRC Press, 2016.

[25] Mark J Van Der Laan and Daniel Rubin. “Targeted maximum likelihood learning”. In: *The international journal of biostatistics* 2.1 (2006).

[26] DH Lawson. “Pharmacoepidemiology: a new discipline.” In: *British medical journal (Clinical research ed.)* 289.6450 (1984), p. 940.

[27] MD Ryuzo Kawamori, MD Yang Wenying, and MD Paul Valensi. “Importance of Observational Studies in Clinical Practice”. In: (2007).

[28] Marlies Noordzij et al. “Study designs in clinical research”. In: *Nephron Clinical Practice* 113.3 (2009), pp. c218–c221.

[29] Sander Greenland and Hal Morgenstern. “Confounding in health research”. In: *Annual review of public health* 22.1 (2001), pp. 189–212.

[30] Samy Suissa and Edeltraut Garbe. “Primer: administrative health databases in observational studies of drug effects—advantages and disadvantages”. In: *Nature clinical practice Rheumatology* 3.12 (2007), pp. 725–732.

[31] Samy Suissa. “Statistical methods in pharmacoepidemiology: principles in managing error”. In: *Drug safety* 6.5 (1991), pp. 381–389.

[32] Ian Shrier et al. “Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles”. In: *American journal of epidemiology* 166.10 (2007), pp. 1203–1209.

[33] Stephen JW Evans. “An agenda for UK clinical pharmacology: Pharmacoepidemiology”. In: *British journal of clinical pharmacology* 73.6 (2012), pp. 973–978.

[34] Alexander T Cohen et al. “Why do we need observational studies of everyday patients in the real-life setting?” In: *European Heart Journal Supplements* 17.suppl_D (2015), pp. D2–D8.

[35] John Concato, Nirav Shah, and Ralph I Horwitz. “Randomized, controlled trials, observational studies, and the hierarchy of research designs”. In: *New England journal of medicine* 342.25 (2000), pp. 1887–1892.

[36] Anton Pottegård. “Core concepts in pharmacoepidemiology: Fundamentals of the cohort and case–control study designs”. In: *Pharmacoepidemiology and Drug Safety* 31.8 (2022), pp. 817–826.

[37] Yasuo Takahashi, Yayoi Nishida, and Satoshi Asai. “Utilization of health care databases for pharmacoepidemiology”. In: *European journal of clinical pharmacology* 68 (2012), pp. 123–129.

[38] Slavenka Janković. *Observational StudiesObservational studies*. Springer Netherlands, 2008.

[39] Cristina Mazzali and Piergiorgio Duca. “Use of administrative data in healthcare research”. In: *Internal and emergency medicine* 10.4 (2015), pp. 517–524.

[40] Suzanne M Cadarette and Lindsay Wong. “An introduction to health care administrative data”. In: *The Canadian journal of hospital pharmacy* 68.3 (2015), p. 232.

[41] Emma H Ulrich et al. “A Review on the Application and Limitations of Administrative Health Care Data for the Study of Acute Kidney Injury Epidemiology and Outcomes in Children”. In: *Frontiers in Pediatrics* 9 (2021), p. 742888.

[42] John G Connolly et al. “Quantifying bias reduction with fixed-duration versus all-available covariate assessment periods”. In: *Pharmacoepidemiology and Drug Safety* 28.5 (2019), pp. 665–670.

[43] Jesper Hallas, David Gaist, and Lars Bjerrum. “The waiting time distribution as a graphical approach to epidemiologic measures of drug utilization”. In: *Epidemiology* (1997), pp. 666–670.

[44] Kyoung-Hoon Kim and Lee-Su Ahn. “A comparative study on comorbidity measurements with lookback period using health insurance database: focused on patients who underwent percutaneous coronary intervention”. In: *Journal of Preventive Medicine and Public Health* 42.4 (2009), pp. 267–273.

[45] Mitchell M Conover et al. “Classifying medical histories in US Medicare beneficiaries using fixed vs all-available look-back approaches”. In: *Pharmacoepidemiology and drug safety* 27.7 (2018), pp. 771–780.

[46] Osemeke U Osokogu et al. “Choice of time period to identify confounders for propensity score matching, affected the estimate: a retrospective cohort study of drug effectiveness in asthmatic children”. In: *Journal of clinical epidemiology* 101 (2018), pp. 107–115.

[47] Jian Sheng Chen et al. “Use of hospitalisation history (lookback) to determine prevalence of chronic diseases: impact on modelling of risk factors for haemorrhage in pregnancy”. In: *BMC medical research methodology* 11.1 (2011), pp. 1–9.

[48] Anders H Riis et al. “Short look-back periods in pharmacoepidemiologic studies of new users of antibiotics and asthma medications introduce severe misclassification”. In: *Pharmacoepidemiology and drug safety* 24.5 (2015), pp. 478–485.

[49] David B Preen et al. “Length of comorbidity lookback period affected regression model performance of administrative health data”. In: *Journal of clinical epidemiology* 59.9 (2006), pp. 940–946.

[50] Mansour TA Sharabiani, Paul Aylin, and Alex Bottle. “Systematic review of comorbidity indices for administrative data”. In: *Medical care* (2012), pp. 1109–1118.

[51] Gerhard Sulo et al. “Effect of the lookback period’s length used to identify incident acute myocardial infarction on the observed trends on incidence rates and survival:

cardiovascular disease in Norway project". In: *Circulation: Cardiovascular Quality and Outcomes* 8.4 (2015), pp. 376–382.

[52] Avram S Bukhbinder et al. "Risk of Alzheimer's disease following influenza vaccination: a claims-based cohort study using propensity score matching". In: *Journal of Alzheimer's Disease* Preprint (2022), pp. 1–14.

[53] Elizabeth A Stuart. "Matching methods for causal inference: A review and a look forward". In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 25.1 (2010), p. 1.

[54] Jill Hardin and Jenna M Reps. "Evaluating the impact of covariate lookback times on performance of patient-level prediction models". In: *BMC Medical Research Methodology* 21.1 (2021), pp. 1–9.

[55] Romain Neugebauer and Mark van der Laan. "Why prefer double robust estimators in causal inference?" In: *Journal of statistical planning and inference* 129.1-2 (2005), pp. 405–426.

[56] Megan S Schuler and Sherri Rose. "Targeted maximum likelihood estimation for causal inference in observational studies". In: *American journal of epidemiology* 185.1 (2017), pp. 65–73.

[57] Menglan Pang et al. "Targeted maximum likelihood estimation for pharmacoepidemiologic research". In: *Epidemiology (Cambridge, Mass.)* 27.4 (2016), p. 570.

[58] Kristin E Porter et al. "The relative performance of targeted maximum likelihood estimators". In: *The international journal of biostatistics* 7.1 (2011).

[59] Mireille E Schnitzer, Erica EM Moodie, and Robert W Platt. "Targeted maximum likelihood estimation for marginal time-dependent treatment effects under density misspecification". In: *Biostatistics* 14.1 (2013), pp. 1–14.

[60] Noémi Kreif et al. "Evaluating treatment effectiveness under model misspecification: a comparison of targeted maximum likelihood estimation with bias-corrected matching". In: *Statistical methods in medical research* 25.5 (2016), pp. 2315–2336.

[61] Ioannis Kosmidis. “Bias in parametric estimation: reduction and useful side-effects”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 6.3 (2014), pp. 185–196.

[62] George Casella and Roger L Berger. “Statistical Inference, Brooks”. In: *Col Publish Co* (1990).

[63] Tony Blakely et al. “Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference”. In: *International journal of epidemiology* 49.6 (2020), pp. 2058–2064.

[64] Miguel Angel Luque-Fernandez et al. “Targeted maximum likelihood estimation for a binary treatment: A tutorial”. In: *Statistics in medicine* 37.16 (2018), pp. 2530–2546.

[65] Sandrine Dudoit and Mark J van der Laan. “Asymptotics of cross-validated risk estimation in estimator selection and performance assessment”. In: *Statistical methodology* 2.2 (2005), pp. 131–154.

[66] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. “Super learner”. In: *Statistical applications in genetics and molecular biology* 6.1 (2007).

[67] K John McConnell and Stephan Lindner. “Estimating treatment effects with machine learning”. In: *Health services research* 54.6 (2019), pp. 1273–1282.

[68] Laura B Balzer and Ted Westling. “Demystifying statistical inference when using machine learning in causal research”. In: *American Journal of Epidemiology* (2021).

[69] Amir Almasi-Hashiani et al. “The causal effect and impact of reproductive factors on breast cancer using super learner and targeted maximum likelihood estimation: a case-control study in Fars Province, Iran”. In: *BMC Public Health* 21.1 (2021), pp. 1–8.

[70] Lisa M Bodnar et al. “Machine learning as a strategy to account for dietary synergy: an illustration based on dietary intake and adverse pregnancy outcomes”. In: *The American journal of clinical nutrition* 111.6 (2020), pp. 1235–1243.

[71] Mireille E Schnitzer, Judith J Lok, and Ronald J Bosch. “Double robust and efficient estimation of a prognostic model for events in the presence of dependent censoring”. In: *Biostatistics* 17.1 (2016), pp. 165–177.

[72] Joshua R Nugent and Laura B Balzer. “A demonstration of Modified Treatment Policies to evaluate shifts in mobility and COVID-19 case rates in US counties”. In: *American Journal of Epidemiology* (2023).

[73] Ronald Herrera et al. “Estimating the causal impact of proximity to gold and copper mines on respiratory diseases in Chilean children: An application of targeted maximum likelihood estimation”. In: *International journal of environmental research and public health* 15.1 (2018), p. 39.

[74] M Alan Brookhart et al. “Variable selection for propensity score models”. In: *American journal of epidemiology* 163.12 (2006), pp. 1149–1156.

[75] Judea Pearl. “Causal inference”. In: *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment-Volume 6*. 2008, pp. 39–58.

[76] Paul W Holland. “Statistics and causal inference”. In: *Journal of the American statistical Association* 81.396 (1986), pp. 945–960.

[77] Donald B Rubin. *Matched sampling for causal effects*. Cambridge University Press, 2006.

[78] Paul R Rosenbaum and Rosenbaum. *Observational Studies*. Springer New York, 2002.

[79] Nicholas A Christakis and Theodore J Iwashyna. “The health impact of health care on families: a matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses”. In: *Social science & medicine* 57.3 (2003), pp. 465–475.

[80] Donald B Rubin. “Estimating causal effects from large data sets using propensity scores”. In: *Annals of internal medicine* 127.8_Part_2 (1997), pp. 757–763.

[81] Donald B Rubin. “Causal inference using potential outcomes: Design, modeling, decisions”. In: *Journal of the American Statistical Association* 100.469 (2005), pp. 322–331.

[82] MA Hernan and J Robins. “Causal Inference: What if. Boca Raton: Chapman & Hill/CRC”. In: (2020).

[83] Donald B Rubin. “Statistics and causal inference: Comment: Which ifs have causal answers”. In: *Journal of the American Statistical Association* 81.396 (1986), pp. 961–962.

[84] Sander Greenland and James M Robins. “Identifiability, exchangeability, and epidemiological confounding”. In: *International journal of epidemiology* 15.3 (1986), pp. 413–419.

[85] Nafisha Lalani, Rachel B Jimenez, and Beow Yeap. “Understanding propensity score analyses”. In: *International Journal of Radiation Oncology, Biology, Physics* 107.3 (2020), pp. 404–407.

[86] Paul R Rosenbaum and Donald B Rubin. “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika* 70.1 (1983), pp. 41–55.

[87] Donald B Rubin. “Estimating causal effects of treatments in randomized and nonrandomized studies.” In: *Journal of educational Psychology* 66.5 (1974), p. 688.

[88] Kara E Rudolph and Elizabeth A Stuart. “Using sensitivity analyses for unobserved confounding to address covariate measurement error in propensity score methods”. In: *American journal of epidemiology* 187.3 (2018), pp. 604–613.

[89] Marco Caliendo and Sabine Kopeinig. “Some practical guidance for the implementation of propensity score matching”. In: *Journal of economic surveys* 22.1 (2008), pp. 31–72.

[90] Jeffrey A Smith and Petra E Todd. “Does matching overcome LaLonde’s critique of nonexperimental estimators?” In: *Journal of econometrics* 125.1-2 (2005), pp. 305–353.

[91] Rajeev H Dehejia and Sadek Wahba. “Propensity score-matching methods for non-experimental causal studies”. In: *Review of Economics and statistics* 84.1 (2002), pp. 151–161.

[92] Onur Baser. “Too much ado about propensity score models? Comparing methods of propensity score matching”. In: *Value in Health* 9.6 (2006), pp. 377–385.

[93] Kara E Rudolph et al. “Optimally combining propensity score subclasses”. In: *Statistics in medicine* 35.27 (2016), pp. 4937–4947.

[94] MILIS Study Group et al. “Digoxin Therapy and Mortality after Myocardial Infarction”. In: *New England Journal of Medicine* 314.5 (1986), pp. 265–271.

[95] Richard A Berk and Phyllis J Newton. “Does arrest really deter wife battery? An effort to replicate the findings of the Minneapolis spouse abuse experiment”. In: *American Sociological Review* (1985), pp. 253–262.

[96] Richard A Berk, Phyllis J Newton, and Sarah Fenstermaker Berk. “What a difference a day makes: An empirical study of the impact of shelters for battered women”. In: *Journal of Marriage and the Family* (1986), pp. 481–490.

[97] William G Cochran. “The effectiveness of adjustment by subclassification in removing bias in observational studies”. In: *Biometrics* (1968), pp. 295–313.

[98] Don Rubin. “Interview with Don Rubin”. In: *Observational Studies* 8.2 (2022), pp. 77–94.

[99] James M Robins. “Causal inference from complex longitudinal data”. In: *Latent variable modeling and applications to causality*. Springer, 1997, pp. 69–117.

[100] James M Robins, Miguel Angel Hernan, and Babette Brumback. *Marginal structural models and causal inference in epidemiology*. 2000.

[101] Stephen R Cole and Miguel A Hernán. “Constructing inverse probability weights for marginal structural models”. In: *American journal of epidemiology* 168.6 (2008), pp. 656–664.

[102] Nicholas C Chesnaye et al. “An introduction to inverse probability of treatment weighting in observational research”. In: *Clinical Kidney Journal* 15.1 (2022), pp. 14–20.

[103] Peter C Austin. “The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies”. In: *Medical decision making* 29.6 (2009), pp. 661–677.

[104] Melissa M Garrido. “Covariate adjustment and propensity score”. In: *Jama* 315.14 (2016), pp. 1521–1522.

[105] Daniel G Horvitz and Donovan J Thompson. “A generalization of sampling without replacement from a finite universe”. In: *Journal of the American statistical Association* 47.260 (1952), pp. 663–685.

[106] Megan S Schuler, Wanghuan Chu, and Donna Coffman. “Propensity score weighting for a continuous exposure with multilevel data”. In: *Health Services and Outcomes Research Methodology* 16.4 (2016), pp. 271–292.

[107] Matthew J Salganik. “Variance estimation, design effects, and sample size calculations for respondent-driven sampling”. In: *Journal of Urban Health* 83.Suppl 1 (2006), p. 98.

[108] William E Barlow. “Robust variance estimation for the case-cohort design”. In: *Biometrics* (1994), pp. 1064–1072.

[109] Wei Pan. “On the robust variance estimator in generalised estimating equations”. In: *Biometrika* 88.3 (2001), pp. 901–906.

[110] Peter C Austin. “Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis”. In: *Statistics in medicine* 35.30 (2016), pp. 5642–5655.

[111] Targeted Learning. “Causal Inference for Observational and Experimental Data”. In: *MR2867111.[Google Scholar]* (2011).

[112] Peter J Bickel et al. *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Springer, 1993.

[113] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. “Estimation of regression coefficients when some regressors are not always observed”. In: *Journal of the American statistical Association* 89.427 (1994), pp. 846–866.

[114] Anastasios A Tsiatis. “Semiparametric theory and missing data”. In: (2006).

[115] Susan Gruber and Mark J van der Laan. “Targeted minimum loss based estimator that outperforms a given estimator”. In: *The International Journal of Biostatistics* 8.1 (2012).

[116] Susan Gruber and Mark Van Der Laan. “tmle: an R package for targeted maximum likelihood estimation”. In: *Journal of Statistical Software* 51 (2012), pp. 1–35.

[117] Mark J van der Laan et al. “Introduction to TMLE”. In: *Targeted Learning: Causal Inference for Observational and Experimental Data* (2011), pp. 67–82.

[118] Cheng Ju et al. “Scalable collaborative targeted learning for large scale and high-dimensional data”. In: *UC Berkeley Division of Biostatistics Working Paper Series. Working Paper* 352 (2016).

[119] Eric C Polley and Mark J Van der Laan. “Super learner in prediction”. In: (2010).

[120] Zhuo Yu and Mark J van der Laan. “Construction of counterfactuals and the G-computation formula”. In: (2002).

[121] Jonathan M Snowden, Sherri Rose, and Kathleen M Mortimer. “Implementation of G-computation on a simulated data set: demonstration of a causal inference technique”. In: *American journal of epidemiology* 173.7 (2011), pp. 731–738.

[122] Tyler J VanderWeele and Stijn Vansteelandt. “A weighting approach to causal effects and additive interaction in case-control studies: marginal structural linear odds models”. In: *American journal of epidemiology* 174.10 (2011), pp. 1197–1203.

[123] Rhian M Daniel et al. “Methods for dealing with time-dependent confounding”. In: *Statistics in medicine* 32.9 (2013), pp. 1584–1618.

[124] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[125] Romain Pirracchio, Maya L Petersen, and Mark Van Der Laan. “Improving propensity score estimators’ robustness to model misspecification using super learner”. In: *American journal of epidemiology* 181.2 (2015), pp. 108–119.

[126] Ron Kohavi et al. “A study of cross-validation and bootstrap for accuracy estimation and model selection”. In: *Ijcai*. Vol. 14. 2. Montreal, Canada. 1995, pp. 1137–1145.

[127] Sudhir Varma and Richard Simon. “Bias in error estimation when using cross-validation for model selection”. In: *BMC bioinformatics* 7.1 (2006), pp. 1–8.

[128] Eric Polley et al. “Package ‘SuperLearner’”. In: *CRAN* (2019).

[129] Terry Therneau et al. “Package ‘rpart’”. In: *Available online: cran. ma. ic. ac. uk/web/packages/rpa pdf (accessed on 20 April 2016)* (2015).

[130] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[131] M Alan Brookhart, Jeremy A Rassen, and Sebastian Schneeweiss. “Instrumental variable methods in comparative safety and effectiveness research”. In: *Pharmacoepidemiology and drug safety* 19.6 (2010), pp. 537–554.

[132] Judea Pearl et al. “Models, reasoning and inference”. In: *Cambridge, UK: CambridgeUniversityPress* 19.2 (2000).

[133] Artur Pokropek. “Introduction to instrumental variables and their application to large-scale assessment data”. In: *Large-scale Assessments in Education* 4 (2016), pp. 1–20.

[134] Ashwin N Ananthakrishnan. “Epidemiology and risk factors for IBD”. In: *Nature reviews Gastroenterology & hepatology* 12.4 (2015), pp. 205–217.

[135] Peter C Austin. “The performance of different propensity-score methods for estimating relative risks”. In: *Journal of clinical epidemiology* 61.6 (2008), pp. 537–545.

[136] Carl De Boor and Carl De Boor. *A practical guide to splines*. Vol. 27. springer-verlag New York, 1978.

[137] Samy Suissa and David A Henry. “The establishment of the Canadian network for observational drug effect studies (cNODES)”. In: *Pharmacoepidemiology and Drug Safety* 20.S1 (2011), S168.

[138] Luis A Garcıa Rodrıguez and Susanne Pérez Gutthann. “Use of the UK general practice research database for pharmacoepidemiology”. In: *British journal of clinical pharmacology* 45.5 (1998), pp. 419–425.

[139] Rohini Mathur et al. “Completeness and usability of ethnicity data in UK-based primary care and hospital databases”. In: *Journal of public health* 36.4 (2014), pp. 684–692.

[140] Achim Wolf et al. “Data resource profile: clinical practice research Datalink (CPRD) aurum”. In: *International journal of epidemiology* 48.6 (2019), 1740–1740g.

[141] Stef Van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate imputation by chained equations in R”. In: *Journal of statistical software* 45 (2011), pp. 1–67.

[142] Marcia P Jimenez et al. “Associations between nature exposure and health: a review of the evidence”. In: *International Journal of Environmental Research and Public Health* 18.9 (2021), p. 4790.

[143] MK Krzyzanowska, R Kaplan, and R Sullivan. “How may clinical research improve healthcare outcomes?” In: *Annals of oncology* 22 (2011), pp. vii10–vii15.

[144] Mark J van der Laan and Susan Gruber. “Collaborative double robust targeted maximum likelihood estimation”. In: *The international journal of biostatistics* 6.1 (2010).

[145] Susan Gruber and Mark J van der Laan. “An application of collaborative targeted maximum likelihood estimation in causal inference and genomics”. In: *The International Journal of Biostatistics* 6.1 (2010).