Machine Learning Methods for Causal Inference with Observational Biomedical Data

Amelia Jean Averitt

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

ABSTRACT

Machine Learning Methods for Causal Inference with Observational Biomedical Data

Amelia Jean Averitt

Causal inference – the process of drawing a conclusion about the impact of an exposure on an outcome – is foundational to biomedicine, where it is used to guide intervention. The current gold-standard approach for causal inference is randomized experimentation, such as randomized controlled trials (RCTs). Yet, randomized experiments, including RCTs, often enforce strict eligibility criteria that impede the generalizability of causal knowledge to the real world. Observational data, such as the electronic health record (EHR), is often regarded as a more representative source from which to generate causal knowledge. However, observational data is non-randomized, and therefore causal estimates from this source are susceptible to bias from confounders. This weakness complicates two central tasks of causal inference: the replication or evaluation of existing causal knowledge and the generation of new causal knowledge. In this dissertation I (i) address the feasibility of observational data to replicate existing causal knowledge and (ii) present new methods for the generation of causal knowledge with observational data, with a focus on the causal tasks of comparing an outcome between two cohorts and the estimation of attributable risks of exposures in a causal system.

Table of Contents

Acknowledgements

This dissertation was made possible by the contributions and support of many people. First and foremost, I would like to thank my advisor, Dr. Adler Perotte. His expertise, guidance, assistance, and patience were instrumental to the completion of this thesis. I had hoped that my PhD would be an opportunity to explore challenging research methodologies that are not readily implemented outside of academia. Dr. Perotte wholeheartedly supported my exploration of these interests, and for this, I am extremely grateful.

I would additionally like to thank my internal committee members – Dr Patrick Ryan and Dr Chunhua Weng. Dr. Ryan helped define the scope of his dissertation and refine the language, and Dr. Weng was the perennial advocate for my work, often reminding me that my research is important and of interest to the scientific community. Thank you to my external committee members, Dr. Jeff Goldsmith and Dr. Rajesh Ranganath, for their generous and helpful suggestions. Thank you all for agreeing to be part of my committee. Without your participation, this work would not have been possible.

In addition to my thesis committee, this work would not have been possible without the contributions of my coauthors, Natnicha (Numfah) Vanitchanant and James (Jimmy) Rogers.

I would like to thank my all of my colleagues over the years at Columbia Department of Biomedical Informatics (DBMI), who – not only helped shaped presentations, papers, and experiments – but were also an incredible source of emotional support. In particular, I would like to thank Michelle Chau, Amanda Moy, Dr. Alexandre Yahi, Dr. Mollie McKillop, Dr. Fernanda Polubriaginof, and Dr. Ben Slovis.

But most of all, I want to thank my friends and family that have always supported me. It is incredibly satisfying to *finally* be able to acknowledge them in the culmination of my work. My mother, Constance Salemi, whose unwavering support and love is the only thing that kept me together through ten years of graduate school and fifteen years in New York City on a shoestring-budget. You are my rock. To my sister, Elizabeth Averitt, who always makes me laugh, no matter the circumstance nor whether or not it's appropriate. To my father, Neil Averitt, who always tells me that he's proud of me. To my best and oldest friends, Erin Tornello and Dr. Julia Brown, who are with me at every milestone. I love you both. To my stepfather, Joe Watson, who always encouraged to pursue the things that scare me. To my stepmother, Kirstin Downey, who is my writing role-model. To my brother, John Averitt, who puts me back in my place. To my niece, Maren Averitt, who brightened up my world during the darkest point of this PhD. And finally to my husband, Evan

Giordanella, who has been with me every day. For every investment I have made into this degree, he has made an equal such investment. I appreciate your patience, tolerance, and support. There are no words to express my gratitude.

I would also like to acknowledge that the novel coronavirus (SARS-CoV-2) made composing this dissertation an absolute nightmare. But mandatory social-distancing may have helped my focus.

Dedication

To my mother, Constance Salemi, and husband, Evan Giordanella.

Chapter 1

Overview of Thesis

Introduction

Statistical inference is the process of determining relationships among entities in a model through the analysis of data. While inference, on its own, refers to the general process of arriving at a logical conclusion based on the validity of former evidence; statistical inference differs in that the reasoning process occurs at the intersection of philosophy, mathematics, and empirical science. These three disciplines contribute to the main roles of statistical inference – the development of algorithms and the inferential arguments which support them (Efron/Hastie 2016). Of the many inference questions posited, the relationship between cause and effect, or *causal inference*, is fundamental. Questions of causal inference are central to many disciplines, such as economics, marketing, and health, but is specific to none. Rather, it is a system that can be widely applied to support causal claims and evaluate their strength.

Causal inference is often explored through the use of a counterfactual – a population that is identical to the treatment arm in all respects, except for the presence of the intervention. However, the counterfactual is never observable, and approximations are needed (Rubin 1986). The current gold-standard for counterfactual causal inference is randomized experimentation. The random allocation of patients is intended to eliminate confounding, as the presence of potentially biasing features should be equal between arms. In this case, the comparator arm then approximates the counterfactual of the treatment arm (Rothman et al. 1998). In the potential outcomes framework, this equality between arms is known as strong ignorability (Rosenbaum/Rubin 1983a). This assumption states that a unit's assignment to a treatment is not a function of that unit's potential outcomes, and that treatment assignment is, therefore, ignorable given their observed features. Causal claims borne from data that satisfy this requirement are regarded as unconfounded as all factors of variation should be equally represented in the treatment and comparator groups (Rubin 1974; Rubin 1986; Rubin 2005). Associations that result from a counterfactual comparison, if found significant by the inference procedure, are recognized as causal knowledge. In biomedicine, causal knowledge is often generated by randomized controlled trials (RCTs), and is later applied in the treatment of patients. This practice of applying causal knowledge is known as evidence based medicine (EBM) (Sackett et al. 1996). The hierarchy of EBM assigns causal knowledge borne from the RCT to the highest level of reputability, as the curated trial population, presence of the control group, and randomized treatment allocation ensure high internal validity (Sackett et al. 1996; Burns et al. 2011). However, randomized experiments, including RCTs, often enforce unrealistic assumptions that impede the generalization of causal knowledge to the real-world (Steckler/McLeroy 2008). More generalizable methods of causal inference would be preferred.

There exist many other methods to support causal inquiry from observational data, including G-estimation (Robins 1986), Pearl's do-calculus (Pearl 1995; Pearl 2000), structural equation modeling (Pearl 2012), Granger causality (Sorenson 2005), and instrumental variables (Becker 2016). However, these methods either require experimental data, assumptions regarding the causal structure, or are intended to model temporal causal structures. In this dissertation, I focus on observational data, making few assumptions about the causal structure, and consider time-varying confounding out of scope.

Problem Statement

Observational data is often lauded as a more externally valid source from which to generate causal knowledge, but it suffers from complexities such as poor quality, irregular sampling, and systematic biases that undermine its use in causal estimation (Hripcsak/Albers 2012; Weiskopf et al. 2013). These weaknesses complicate two central tasks of causal inference - the replication or evaluation of *existing* causal knowledge, and the generation of *new* causal knowledge.

When evaluating *existing* causal knowledge through replication, the greatest obstacles in the use of observational data are the effective management of observed and unobserved confounding and the insufficient reporting of the experimental population. Even with precise definitions of observational populations, it remains difficult to truly replicate causal estimates as all unit-level features of variation are typically not disclosed and the majority of causal estimates are reported in summary. This has the additional consequence of making the extent of generalizability unknown. This failure of an RCT to represent an indicated population may be the result of cohort selection for a single condition of interest that is further narrowed by a breadth of eligibility criteria. This final cohort population may stand in stark comparison to the intended target population (Wales 2009). However, the appropriate application of causal knowledge would require that the intended real-world population be the same as the experimental population, but this is a challenging task given that the RCT-reported population characteristics are limited and only presented very coarsely.

The confounding typical of observational data may also interfere with popular methods of generating *new* causal knowledge. When comparing two cohorts, the calculation of unbiased causal estimates from this imperfect data source is often framed as identifying a *natural experiment.* Natural experiments are a type of observational study in which researchers do not have the ability to assign the treatment, but treatments are nonetheless assigned nearly randomly. They are most valid when they closely resemble a true experimental setting, in which treatment is randomized (Meyer 1995; Shadish et al. 2002; Academy of Medical Sciences 2007; Craig et al. 2012). Popular pre-analysis methods for approximating natural experiments include matching, in which treatment units are paired with similar comparator units based on the pre-treatment features (Wilks 1932; Cochran 1953; Greenberg 1953; Billewicz 1965; Rubin 1973a); and weighting, in which units are disproportionally considered so that the weighted expectation of features are similar across arms (Czajka et al. 1992; Robins et al. 2000; Lunceford/Davidian 2004). All weighting methods generalize matching methods, and conversely, all types of matching are special cases with discrete weights (Imai 2013). Under a matching procedure, units may go unpaired, which is inefficient and may introduce new bias (Rosenbaum/Rubin 1985b; King 2011b). Weighting is a more efficient method for identifying a natural experiment. However, under many weighting techniques, downstream estimates may be unstable. In this dissertation, I present a stable weighting method to support causal inference from two cohorts of observational data. This method will increase the confidence of causal claims from observational data and may permit the identification of effective interventions and improve outcomes.

New causal knowledge is also generated through attributable risk (AR) estimation. ARs are the proportion of an outcome in a population that could be prevented by elimination of a causal exposure from the population if there are (i) no interactions between causal exposures and (ii) all other effects of exposures are removed (Levin 1953). Typically, estimation of ARs would be based upon knowledge of the relevant causal graph. However, in the setting of many potential exposures – the *high-dimensional setting* – causal graph construction may be impractical. To estimate the ARs of many exposures simultaneously in the absence of the causal graph, model specification becomes increasingly important. In this dissertation, I explore a particular the specification for AR estimation from unstructured binary exposures and outcomes. In the absence of the causal graph, typical methods of high-throughput

AR estimation include the Gamma Poisson Shrinker (GPS) (Bate/Evans 2009) and Penalized Logistic Regression (PLR) (Hahn et al. 2017). However, GPS is univariate and PLR is an unlikely causal model. Additionally, although techniques that support adjustment for confounding exist for these methods, these models may still be subject to bias by non-causal pathways between the exposure and the outcome, known as backdoor paths (Pearl 1995). In a backdoor path, variables are not statistically independent, but one does not necessarily cause the other. In this situation there are two sources of association between the exposure and the outcome; (i) the true causal effect of the exposure on the outcome, and (ii) the non-causal effect through the backdoor path (Blackwell 2013). Consider the canonical example of lighters (L), smoking (S), and lung cancer (LC). Though lighters may appear to be causally related to lung cancer (L $\rightarrow$ LC), this apparent association is driven via a backdoor path of the relationships between smoking and lighter (S $\rightarrow$ L), and smoking and lung cancer (S $\rightarrow$ LC). This confounding via the backdoor path obscures the lack of a true causal association (L $\rightarrow$ LC). When analyzing observational data, such a backdoor path may emerge through collinear variables with causal associations with exposure and outcome; in this case between lung cancer and lighters. Developing methods that correct for such collinearities, can be leveraged to resolve the underlying causal relationships we seek.

I present a multivariate, latent variable model for AR estimation with observational data that effectively handles observed confounding, collinearities, and may be more robust to unobserved confounding via backdoor paths. Furthermore, unlike comparator methods, the method I propose supports the evaluation of global risks, the prediction outcomes, and estimation of causes at the individual level.

Purpose of the Study

This dissertation presents research which investigates (i) the replication and evaluation of existing causal knowledge with observational data and (ii) the generation of new causal knowledge with observational data.

In Aim 1 of this dissertation, I seek to determine the extent to which observational data can be used to replicate *existing causal knowledge* from randomized experimental results. If residual bias is present, we seek to characterize and quantify this bias, to better understand the limitations of unmanipulated observational data in a causal inference setting. In Aims 2 and 3 of this dissertation, I seek to develop methods that support *generating new causal knowledge* from this biased data source. Aim 2 will address reducing bias in causal estimates from observational data, when confronted with the inference task of comparing two cohorts; and Aim 3 will seek to produce unbiased, high-throughput attributable risk estimates.

Research questions, hypotheses, specific aims, and experimental designs can be found in the sections that follow.

Research Questions and Hypotheses

Aim 1. Determine the feasibility of observational data to replicate existing causal knowledge.

Aim 1.1 *Assess the ability to replicate causal claims from RCTs using electronic health record (EHR) data.*

Research Question	*Does the construction of observational cohorts according to RCT eligibility criteria encourage the causal effect estimate to converge with that reported in the trial?*
Hypothesis	Eligibility criteria are not sufficient to identify an observational population in which experimental causal estimates will replicate.

Aim 1.2 *Examine potential sources of residual bias in effect estimates from observational data sources.*

Research Question	*Why do observational causal effect estimates fail to converge with that reported in the RCT?*
Hypothesis	The residual bias between observational effect estimates and RCT effect estimates is due to distributional differences in potentially confounding variables.

Aim 2. Develop a method to identify natural experiments within observational data.

Aim 2.1 *Implement a generative adversarial network, Counterfactual χ-GAN, to learn balancing weights.*

Research Question	*Can a generative adversarial network (GAN) be leveraged to learn feature-balancing weights?*
Hypothesis	In simulation, a GAN-based model will identify units that are generated from the same underlying distribution, and assign these units greater weights, thereby improving feature balance.

Aim 2.2 *Apply the Counterfactual χ-GAN to observational datasets.*

Research Question	*Can a generative adversarial network (GAN)-based model improve feature-balance for noisy observational cohorts?*
Hypothesis	The cGAN model will learn feature-balancing weights for two cohorts; the resulting weighted metrics will be more similar to that reported in truly randomized trials.

Aim 3. Develop a method for high-throughput causal attributable risk estimation with observational data.

Aim 3.1 *Implement a probabilistic model, Noisy-Or Risk Allocation (NORA), and develop efficient probabilistic inference procedures.*

Research Question	*Will a Bayesian model that encodes the assumption of causal independence produce attributable risk estimates that are less biased than other state-of-the-art methods?*
Hypothesis	In simulation, a Bayesian model that encodes the assumption of causal independence will be less biased from the ground truth than Logistic Regression.

Aim 3.2 *Apply the NORA model to observational datasets.*

Research Question	*Can a Bayesian model that encodes the assumption of causal independence be used to support high-throughput attributable risk estimation with observational data?*
Hypothesis	The attributable risk estimates learned by NORA will coincide with causal relationships that are acknowledged in biomedical literature.

Experimental Design Associated with Hypotheses

Aim 1

Aim 1.1 Extract inclusion and exclusion from a published RCT protocol and literature. Incrementally add criterion to an observational cohort that is constructed according to the indication of the RCT. This observational cohort is queried multiple times, with each query being subject to the addition of new eligibility criteria of the target RCT. With each incremental criterion (i) the unadjusted odds ratio (OR) for an endpoint and (ii) the OR adjusted by matching, are calculated and compared to the OR that was calculated for this endpoint at the RCTs close.

Aim 1.2 Using the Observational Health Data Science and Informatics (OHDSI) ATLAS tools, curate observational cohorts to match RCT populations. The trial's indication will identify a core set of patients to which inclusion and exclusion criteria are applied to appropriately narrow the cohort. Subjects who remained eligible after this pruning stage comprise the observational cohort of interest. This cohort is then queried to contrast the characteristics of the observational cohort with the RCTs participant features that are reported in the *Demographics Table* ("Table 1").

Aim 2

Aim 2.1 Design and implement a model based on a generative adversarial network (GAN), the Counterfactual χ-GAN (cGAN), that will learn feature-balancing weights for two cohorts. To determine the correctness of implementation, the cGAN and comparators will be applied to simulated data of two cohorts, for which a portion of each cohort is generated from the same distribution. The units that arise from this same distribution will be more similar in their features and should have higher weights. Outcomes will be simulated according to the distribution of origin to accommodate analysis of the biasedness of the average treatment effects (ATE).

Aim 2.2 Apply the Counterfactual χ-GAN (cGAN) and a variety of other comparator methods to observational cohorts from electronic health record (EHR) data, constructed according to RCT indications. The successful application to real-world clinical data will be evaluated by feature balance between the cohorts; measured by the absolute standardized difference in the means (ASDM).

Aim 3

Aim 3.1 Develop a Bayesian, probabilistic model, the Noisy-Or Risk Allocation model (NORA), for the estimation of ARs from observational data. NORA considers multiple independent causes in a setting with binary exposures and outcomes. To assess the correctness of implementation, NORA and logistic regression (LR) are applied to simulated causal system, in which unobserved confounding exists. NORA and LR are tasked with estimating the AR for a single exposure in our simulated data. Estimates from NORA and LR are compared to the ground truth to evaluate robustness to confounding.

Aim 3.2 Apply NORA to observational datasets from the NewYork-Presbyterian Hospital electronic health record (EHR). Exposure-outcome datasets are curated according to several causal relationships acknowledged in published literature. NORA and comparator methods – including direct calculation using the Levin 1953 attributable risk definition; approximation by disproportionality methods; and regression-based methods – are applied to these datasets to estimate the ARs of the exposures. NORA and comparators are evaluated on (i) their ability to make AR estimations that coincide with the literature; (ii) their predictive power for an individuals outcome; and (iii) their ability to rationalize over an individuals exposures.

Significance

The methods put forth in this dissertation will help to improve causal inferences from observational data sources.

The ability to evaluate and replicate causal relationships from observational data may confer many benefits. It may promote a more principled and data-driven practice of evidence-based medicine, and encourage the current system of evidence-generation to provide causal estimates from a more representative population. These actions may contribute to more generalizable treatment effects that may be easily validated or replicated with observational data.

The machine-learning methods to generate new causal knowledge from observational data sources advance both the computer science and clinical communities. The methods presented in this dissertation are novel contributions to the breadth of machine learning models that already exist. Future research may build upon this dissertation's models to further improve causal inferences from observational data. The ability to generate new causal knowledge – by comparing two cohorts or AR estimation – is an improvement over the current standard of inferences from experimental design. The methods I present provide efficient and accurate inferences, which will aid the clinical community to get better care to patients faster.

Questions of causal inference are broadly applicable to domains in which there is actionable uncertainty, and may therefore benefit a wide audience. All of methods within this dissertation may permit researchers across domains and institutions to optimize outcomes with interventions.

Contributions

Aim 1. A number of techniques exist to improve and support causal estimates from observational data, but at present, there is no widely-used framework to evaluate modeling assumptions relative to experimental data. RCTs, which we accept to be the least biased source of causal knowledge, can be compared to estimates generated from observational data and, thus, provide a methodology to assess the validity of causal claims and a platform with which to evaluate inference methods.

Aim 2. The cGAN is an effective method of learning feature-balancing weights to support counterfactual inference between two cohorts. The application of the model to real-world data could provide an alternative means to causal inference from observational data. Furthermore, if we assume that all potentially confounding variables are observed and included as features, a superiority of cGAN in learning balancing weights, suggests that average treatment effects (ATE) borne from cGAN-weighted cohorts would be less biased than those estimates generated from typical weighting methods.

Aim 3. NORA offers advantages over traditional methods of AR estimation. Unlike comparator methods, NORA supports both local and global inferences. The likelihood of the model encodes an intuitive and simplifying assumption of causality. In the absence of causal graph construction, such simplifying assumptions can be powerful tools for estimation even in the context where the assumptions are only partially met. NORA is able to scale to very high dimensions due to the inherent regularizing effect; and early simulations suggest that the model is robust to unobserved confounding and collinearity. Our results show that NORA (i) predicts outcomes with similar or better performance than related methods, (ii) recovers known, clinically meaningful AR estimates, and (iii) produces interpretable estimates of the causes for an individual's outcome.

Chapter 2

Background and Related Work

Historical Background

EBM and Experimental Data Since its inception in the 1990s, Evidence Based Medicine (EBM) has become the standard of operation for clinicians. The practice advocates a framework of clinical care which optimizes patient health through the judicious consideration and application of medical evidence (Djulbegovic/Guyatt 1976; Djulbegovic/Guyatt 2017a; Djulbegovic et al. 2009; Sackett et al. 1996). At the core of the EBM philosophy is the credibility of the medical evidence – the application of any medical evidence is only justifiable if we first believe that evidence is credible (Djulbegovic/Guyatt 1976). EBM encourages clinicians to seek the most reputable evidence according to a hierarchy of study quality (Sackett et al. 1996). The quality of a study is characterized by both the internal and external validity; where internal validity refers to the extent to which a causal conclusion is warranted and is typically measured by the absence of systematic error, and external validity refers to the extent to which causal relationships of a study are able to persist over variation in persons and treatment settings (Campbell/Stanley 1963). The hierarchy of EBM assigns the randomized controlled trial (RCT) to the highest level of reputability, as the curated trial population, presence of the control group, and randomized treatment allocation ensures the highest internal validity (Burns et al. 2011).

It is often asserted that internal validity is a prerequisite for external validity, as study results that fail to capture the true effect due to bias necessarily cannot be expected to generalize outside of the study population (Calder et al. 1982; Dekkers et al. 2010; Higgins et al. 2011). As such industry, academia, and regulatory and government agencies alike, put high importance on the internal validity of trials, prioritizing scientific rigor of the experiment over its ability to replicate outside of the trial setting (Higgins et al. 2011). Similarly, most empirical assessments of study quality for EBM, emphasize the evaluation of internal validity, by limiting the scope of their assessment to factors such as randomization, allocation, blinding, follow-up, and attrition (Moher et al. 1996; Moher et al. 2010; Harbour/Miller 2001; Dijkers 2013). A common design element to support high internal validity of a trial are selection criteria or eligibility criteria. If we consider trial patients to be a function of their features, such as demographics, laboratory measurements, and medical history; the eligibility criteria define the features that all patients in a study must share. In theory, the eligibility criteria support internal validity by ensuring the homogeneity of the study population and reducing confounding. As such, eligibility criteria may increase the prospect of uncovering the true association between an intervention and outcome (Velasco 2010).

When operationalized, the eligibility criteria are represented as inclusion and exclusion criteria (Campbell/Stanley 1963; Hyman 1982; Anderson-Cook 2005). With every addition of a criterion to a study population, it results a different sub-population and increasingly controlled conditions (Velasco 2010). These restrictive criteria may afford high internal validity, but it often comes at the expense of external validity and, consequently, the successful practice of EBM (Rothwell 2006).

Because RCT's are designed to support high internal validity and as such, eligibility criteria are employed with the express purpose to construct a homogeneous, predictable cohort; these trials may fail to incorporate patients with characteristics that yield more variation in the treatment effect and may be more representative of the real world. This may result in making any trial evidence, such as the average treatment effect (ATE), poorly generalizable (Moher et al. 1996; Britton et al. 1999; Wales 2009; Karanis et al. 2016; Stuart et al. 2015). However, is fundamentally at odds with the practice of EBM. Because underlying this practice, is the assumption that an intervention will show a similar effect in a real-world population as shown in the experimental, RCT population. As such, there is an inherent mismatch in what EBM acknowledges to be credible medical evidence and what evidence may support replicable treatment effects in the real-world.

The relationship between RCTs and EBM may also be viewed in another manner. A mismatch between the trial population and the real-world can be simply distilled into the fact that the distributions of subpopulations that result from the application of the eligibility criteria – may be unequal. Because causal effects are not homogeneous, each subpopulation may respond differently. This is known as the *heterogeneity of treatment effect* (Longford 1999; Kravitz et al. 2004; Gabler et al. 2009). The varied causal responses from these subpopulations are then aggregated into a single metric, which is the treatment effect that is clinicians expect will replicate. A failure to consider all of these subpopulations when randomizing in the experimental setting may result in bias (Lachin et al. 1988; Kernan et al. 1999) or poor external validity (Illari et al. 2011).

EBM and Non-Experimental Data To have a fully generalizable effect estimate from experimental data would require (i) that the treatment effect be the same across all subpopulations, or (ii) that you have a priori knowledge of all subpopulations and are able to randomize across them. These are both infeasible. Alternatively, we could use observational data to create causal knowledge and support EBM.

Observational data is data that is passively collected, ”without making any engineering adjustments to the collection process beyond those adjustments that are part of a normal operation” (Czitrom 1997). Importantly, electronic health records (EHR) is a type of observational data (Murdoch/Detsky 2013). Since the enactment of Health Information Technology for Economic and Clinical Health (HITECH) in 1997, the collection of electronically formatted clinical data has greatly increased (Charles 2013). This increase in observational data stores has similarly occurred in other countries with EHR mandates (Heinze et al. 2011; Faxvaag et al. 2011; Tejero/Torre 2012; Shah 2012; Mense et al. 2013).

Observational data offers many advantages over experimental data. First and foremost, there is a lot of it, which provides new opportunities and avenues for pattern recognition (Imai et al. 2009). Observational data is suitable for studying rare outcomes, which is often not supported in the experimental setting. But most importantly, observational data is *representative*. That is, it is more likely to include a broader representation of the at-risk population, which makes inferences from this data source potentially more externally valid (Concato 2004; Thadhani 2006; Kleinberg/Hripcsak 2011). If using observational data, we are much more likely to capture all subpopulations of interest and produce a causal effect estimate that is externally valid.

Observational data is more representative and can yield more generalizable estimates, but observational data itself is plagued with problems that make these estimates unreliable. Observational data is inaccurate, complex (Hripcsak/Albers 2012), incomplete, especially with regards to poor documentation, breadth, and predictive power of the data (Weiskopf et al. 2013). But most importantly, observational data is biased. The lack of randomization, that is always present in the RCT, renders studies with observational data susceptible to biases (Höfler 2005; Dahabreh et al. 2012).

Herein is the central trade off with observational data – it's use confers external validity but reduces internal validity. EBM could greatly benefit from use of this data, but confidence in estimates from this source require that we improve the internal validity by addressing the confounders that arise from the lack of randomization.

Review of Theories of Causal Inference

Causal inference refers to the process of drawing a conclusion about cause and effect relationships (Vogt/Johnson 2011). Research on causal inference comes from a variety of disciplines (Stuart 2010) statistics (Holland 1986; Thrusfield 2017; Rubin 2012) epidemiology (Rothman 2000; Brookhart et al. 2006) sociology (Morgan/Harding 2006) political science (Ho et al. 2007) social science (Sobel 2000).

Epistemological Frameworks. In order to rectify experimental and observational causal inference, we should identify a single theoretical philosophy. This process is known as identification strategy (Angrist et al. 1996; Angrist/Pischke 2010).

INUS Conditons. INUS conditions, sometimes referred to as *Necessary and Sufficient Conditions* are a tool to aide in the search for precise definitions of conditions that must be met in order for a phenomena to be truly present. This framework asserts the four following definitions for defining a precise causal relationships between an exposure (x) and an outcome (y).

> Necessary causes. *for x to be a necessary cause of y, then the presence of y necessarily implies the prior occurrence of x. The presence of x, however, does*

not imply that y will occur.

Sufficient causes. *for x to be a sufficient cause of y, then the presence of x necessarily implies the subsequent occurrence of y. However, another cause z may alternatively cause y. Thus the presence of y does not imply the prior occurrence of x.*

Contributory causes. *for x to be a contributory cause of y, then the presence of x permits the presence of y, but not with certainty. A contributory cause may be neither necessary nor sufficient but it must contribute to the presence of the outcome* (iSTAR Assessment 2011).

J. L. Mackie proposed that true causes are at a minimum *INUS conditions* – ”Insufficient but Necessary parts of a condition which is itself Unnecessary but Sufficient.” Mackie notes that outcomes often have a ”plurality of causes” (p.61); a consequence of which is that an outcome may be resultant from more than one distinct exposures (Mackie 1965).

Criteria-Based. I refer to the frameworks that enumerate principles that must be present for a causal relationship to exist, as *criteria-based.*

Hill Criteria

The Hill Criteria draws upon biological and environmental axioms with scientific knowledge to put forth what he describes as ”aspects of association” that capture the notion that some causal effects are plausible given evidence and others are not. The criteria include (i) strength of association (ii) consistency (iii) specificity

(iv) temporality (v) biological gradient (vi) plausibility (vii) coherence (viii) experiment and (xi) analogy. Of these, the only criteria that *must* be present for a potential causal relationship to exist is temporality, Popular in epidemiology, these criteria explicitly support drawing causal conclusions of observational data sources, which involved disentangling true causal effects from biases (Hill 1965).

Susser Criteria

Susser's Criteria is very similar to Hill's, but slightly smaller in dimension. The Susser Criteria include (i) association (ii) time order (iii) direction (iv) strength (v) specificity (vi) consistency (vii) predictive performance and (viii) performance. Unlike the Hill Criteria, the Susser criteria require that three of the criteria must be present in order for a causal relationship to exist – association, time order, and direction (Susser 1973).

Probabilistic Causality. Probabilistic causality is a framework which characterizes the relationship between cause and effect using probability theory (Good/Suppes 1972; Salmon 1988; Pearl 1995; Yu et al. 2010). The fundamental concept behind this framework is that exposures that are true causes alter the probabilities of outcomes. However, like necessary causes, the outcome may still occur in the absence of an exposure, or fail to occur in the presence of an exposure. Under this framework, true causal exposures (E) will increase the probability of an outcome (O). This is often expressed by conditional probabilities, and is formalized by the following simple expression;

$$P(O|E) > P(O) \tag{2.1}$$

Singular Causation. A subset of probabilistic causality is Singular Causation. Where the above framework of probabilistic causality addresses *general* or type-level inquiries, Singular Causation addresses *singular, token level*, or *actual* inquiries (Hausman 2005). This framework largely differentiates itself from the general inquiry in that it considers how probabilities change over time (Ray 1992; Kvart 2004).

Counterfactual Reasoning. The preferred framework of causal inference in biomedicine, Counterfactual Reasoning leverages if-clauses that are contrary to the truth to investigate what would have happened had the world been different (Lyon 1967; Lewis 1973a, 1973b; Mackie 1980). Counterfactual conditionals are often framed in the statement, *if E had not occurred, O would not have occurred* (Goodman 1947).

The counterfactual conceptualization of causal inference was first popularized in the experimental setting (Neyman 1923; Fisher 1935; Neyman et al. 1935; Cochran/Cox 1950; Payne 2015; Cox 1958; Winship/Morgan 1999). The theory was later formalized to the non-experimental setting (Rubin 1974, 1977, 1978; Rubin 1980; Rubin 1986; Rubin 1990; Pratt/Schlaifer 1984). For this suitability with observational data, this will be the preferred framework used throughout the proposal. For this reason, I will briefly discuss the theoretical details.

Potential Outcomes. Consider a causal setting in which we are interested in the

	$\boldsymbol{T=0}$	$\boldsymbol{T=1}$
$\mathbf{Y_0}$	$Y_0^{T=0}$	$Y_0^{T=1}$
$\mathbf{Y_1}$	$Y_1^{T=0}$	$Y_1^{T=1}$

Table 2.1: Potential Outcomes by State Assignment

binary outcome, Y. For any outcome of interest, we begin by assuming that units may be a priori exposed to two *states* (T) but that each unit is ultimately only exposed to one state. I herein refer to these states as *treatment* ($T = 1$) and *control* ($T = 0$). Both the treatment and control state are characterized by a set of unseen conditions. A unit's exposure to the states' conditions will dictate their development of the outcome of interest.

Foundational to this framework, is the assumption that units have *potential outcomes*, Y_i in both states. Regardless of which state the unit is exposed to, units have potential outcomes for (i) the state in which they are exposed and, (ii) the states in which they are not exposed. Note that in practice, only the potential outcomes for the state in which the unit was exposed is observable. Units assigned to the treatment group, $T = 1$, only have observable outcomes in the outcome state, Y_1; and conversely units assigned to the control group, $T = 0$, only have observable outcomes in the outcome state, Y_0. Referring to Table 2 the potential outcomes, $Y_1(T = 0)$ and $Y_0(T = 1)$ are unobservable. Simply, the counterfactual framework asserts that individuals have potential outcomes in all states, though they can only be observed in one state (Winship/Morgan 1999).

Causal Effects. Throughout causal inference, we are interested in quantifying the extent to which the treatment state affects the outcome, as compared to the control state. Using the notation from (Winship/Morgan 1999), for each individual i, the causal effect is calculated as the difference between the two potential outcomes in the treatment and control states. This is known as the Individual Treatment Effect (ITE) (Equation 2.2).

$$ITE_i = Y_{1,i} - Y_{0,i} \tag{2.2}$$

When summarized over a large number of units, this quantity is averaged, and is known as the unconditional Average Treatment Effect (ATE) (Equation 2.3).

$$\begin{aligned} ATE &= \mathbb{E}[Y_1 - Y_0] \\ &= \mathbb{E}[Y_1] - \mathbb{E}[Y_0] \end{aligned} \tag{2.3}$$

When estimating the ATE ($A\hat{T}E$) from a sample of observational data in which units' behaviors are not known, the naive estimate is given by the difference between the sample mean of the outcome of units in treatment state and the sample mean of the outcome for units in the control state. Equation 2.4.

$$A\hat{T}E_{naive} = \mathbb{E}\big[Y_1|T=1\big] - \mathbb{E}\big[Y_0|T=0\big] \tag{2.4}$$

Each term of the naive $A\hat{T}E$ is a conditional corresponding to a treatment state. The conditional ATE for those assigned to the treatment state ($T = 1$) is given by Equation 2.5, and is often called the *average treatment effect of the treated* or ATT. And for those assigned to the control state ($T = 0$), the conditional ATE is given by

Equation 2.6 and is called the *average treatment effect of the control*, or ATC.

$$\mathbb{E}[ATE|T=1] = \mathbb{E}\left[Y_1^{T=1} - Y_0^{T=1}\right] \tag{2.5}$$

$$\mathbb{E}[ATE|T=0] = \mathbb{E}\left[Y_1^{T=0} - Y_0^{T=0}\right] \tag{2.6}$$

In observational studies, the naive $A\hat{T}E$ estimator (2.4) will only converge with the true, population ATE (2.3) if both the proportion of units assigned to the treatment state (π) and the four potential outcomes (Table 2) are known. A decomposition of this relationship, using the definitions of ATT and ATC, is shown in Equation 2.7.

$$\begin{aligned} ATE =& \pi(Y_1|T=1) + (1-\pi)(Y_0|T=0) \\ =& \pi\left[\mathbb{E}[Y_1^{T=1}] - \mathbb{E}[Y_0^{T=1}]\right] + (1-\pi)\left[\mathbb{E}[Y_1^{T=0}] - \mathbb{E}[Y_0^{T=0}]\right] \\ =& \left[(\pi)\mathbb{E}[Y_1^{T=1}] + (1-\pi)\mathbb{E}[Y_1^{T=0}]\right] - \left[(\pi)\mathbb{E}[Y_0^{T=1}] + (1-\pi)\mathbb{E}[Y_0^{T=0}]\right] \end{aligned} \tag{2.7}$$

Both $Y_1^{T=0}$ and $Y_0^{T=1}$ exist in theory, but recall that only one potential outcome is ever observable for a single unit. This is widely called *the fundamental problem of causal inference.* This inhibits the direct calculation of ITE or ATE. In order for the naive estimator (Equation 2.4) to be unbiased, it is sufficient that $\mathbb{E}[Y_1^{T=1}] = \mathbb{E}[Y_1^{T=0}]$ and $\mathbb{E}[Y_0^{T=0}] = \mathbb{E}[Y_0^{T=1}]$. Substituting these equivalencies in the ATE (Equation 2.8) recovers the naive estimator in Equation 2.4, when this assumption is met. This highlights the requirements for unbiased ATE estimation.

$$\begin{aligned} ATE =& \big[(\pi)\mathbb{E}[Y_1^{T=1}] + (1-\pi)\mathbb{E}[Y_1^{T=1}]\big] - \big[(\pi)\mathbb{E}[Y_0^{T=0}] + (1-\pi)\mathbb{E}[Y_0^{T=0}]\big] \\ =& \mathbb{E}[Y_1^{T=1}]\big[(\pi + (1-\pi))\big] - \mathbb{E}[Y_0^{T=0}]\big[(\pi + (1-\pi))\big] \\ =& \mathbb{E}[Y_1^{T=1}] - \mathbb{E}[Y_0^{T=0}] \end{aligned} \tag{2.8}$$

Assumptions. To ensure that the proxies for unseen potential outcomes yield unbiased causal effect estimates, we enforce the assumptions of the counterfactual framework.

SUTVA. Stable Unit Treatment Value Assignment (SUTVA) requires that ”the [potential outcome of] one unit should be unaffected by the particular assignment of [states] to the other units” (Cox 1958; Rubin 1986). This is colloquially known as *non-interference between states* (Rubin 1980; Rubin 1986; Rubin 1990). The canonical example of a SUTVA violation is from agriculture. Consider we wanted to measure the causal effect of a fertilizer on crop yield. We treat some plots with the fertilizer and the others are non-treated. If a heavy rain falls and the run-off from the treated plots flows into the untreated plots, then the outcome of the untreated plots is no longer a function of the state to which it was assigned.

In addition to interference by states, SUTVA may also be violated when there are hidden states (Laffers/Mellace 2016) or when the states alter the potential outcomes (Garfinkel et al. 1992). Generally, SUTVA cannot be confirmed by the data, and is assumed true unless there is explicit violation.

Ignorability. This assumption states that a unit's assignment to a state is not a function of that unit's potential outcomes. More simply, state assignment is unconfounded, and therefore ignorable (ROSENBAUM/RUBIN 1983a; ROSENBAUM/RUBIN 1983b; ROSENBAUM/RUBIN 1984). In the epidemiology literature, this is referred to as *exchangeability.* In observational studies, causal effects are generally non-estimable because units in the treatment state and the control state are not exchangeable. If this assumption is met, the treatment and control groups are identical insofar as the potential outcomes are concerned (VANDERWEELE/HERNÁN 2013).

For the purposes of this proposal, I will make the distinction between the the conditional and unconditional forms of this assumption.

Unconditional Ignorability Both of the potential outcomes, Y_1 and Y_0, must be jointly independent of the assigned state. Equation 2.9 (RUBIN 1974). When the researcher has the ability to manipulate state assignment, as is the case in perfect randomization, this form of ignorability is satisfied.

$$Y_0, Y_1 \perp T \tag{2.9}$$

Conditional Ignorability In observational studies, when the researcher does not have the ability to alter state assignment, conditional ignorability (Equation 2.10) must be satisfied. Under this condition, the potential outcomes, Y_1 and Y_0, must be conditionally independent of state assignment given the value of their features, X.

$$Y_0, Y_1 \perp T \quad | \quad X \tag{2.10}$$

When the conditional form of ignorability is upheld, the ATE can be estimated through stratification (Winship/Morgan 1999). If the set of stratifying features, S, accounts for all confounding between the treatment state and the control state, then the unconditional ATE can be estimated as a weighted sum of the outcome over strata.

$$\sum_{S} \Big[\mathbb{E}[Y|T=1, S=X] - \mathbb{E}[Y|T=0, S=X] \Big] Pr(S=X) \tag{2.11}$$

In both the conditional and unconditional forms of ignorability, satisfaction of this assumption has the important consequences of making the features in both the treatment and the control states, distributionally equal. This equality of the empirical distribution, $\tilde{F}(\cdot)$, is summarized by, Equation 2.12

$$\tilde{F}(X|T=1) = \tilde{F}(X|T=0) \tag{2.12}$$

As proved by (Rosenbaum/Rubin 1983b) a causal effect can be accurately estimated from observational data so long as the aforementioned assumptions hold and the units can be compared according to their features or function thereof.

Review of Methods of Causal Inference with Observational Data

The Comparison of Two Cohorts

Using the theory of counterfactual inference, we can design an observational study such that strong ignorability is upheld and the ATE will be less biased (Imai et al. 2009). These designs can broadly be broken into two approaches – The first are *pre-analysis* manipulations, such as matching and weighting, which create a pseudo-population with feature-balance. The second are *peri-analysis* manipulations, such as statistical adjustment, which remove the effect of specified confounders. Other methods for causal estimation with observational data exist, such as graphical models & Bayesian networks, instrumental variables, structural equation modeling, and Granger causality, but they are considered out of scope for this review.

Matching

Matching is the most developed and popular strategy for causal analysis in observational studies (Pearl 2010; King/Nielsen 2018). In practice, this involves pre-processing the data such that each treatment unit is paired with a similar control unit based on the pre-treatment features, (X). In line with the assumption of strong ignorability, the goal of matching is to achieve feature-balance (Ho et al. 2007; Imbens 2009; Morgan/Winship 2007). In this condition, important factors of variation are equal between treatment arms, and the only difference that remains is the presence or absence of the treatment (Greenland et al. 1999; Zhao 2004). In line with this end-result, all matching methods couple units from the treatment group ($T = 1$) with units from the control group ($T = 0$) that share similar or exact values of observable features (Imai et al. 2009; Zubizarreta 2012; Papanicolas/Smith 2014; Nielsen 2016). The matching methods reduces bias in average treatment effect (ATE) estimate (Wilks 1932; Cochran 1953; Greenberg 1953; Billewicz 1965; Rubin 1973a). There are a number of different kinds of methods of matching, but each matching procedure is a combination of a (i) balancing metric, (ii) distance measure, and (iii) matching procedure.

> **Balancing Metric.** The metric or metrics that we want to match units on. A balancing metric $b(x)$ is a function of the observed features such that the conditional distribution of x given $b(x)$ is the same for the treated and comparison groups. Equation 2.13 (Rosenbaum/Rubin 1983b; Zhao 2004). Balancing Metrics can be grouped into (i) Raw Features and (ii) Balancing Scores.
>
> $$\tilde{F}(X|T = 1, b(x)) = \tilde{F}(X|T = 0, b(x)) \tag{2.13}$$
>
> *i. Raw Features.* are a fine balancing metric. It requires matching each treatment unit to a control unit with exactly the same values on all features, and then disregarding all non-matched units (Rosenbaum et al. 2007; Imai et al. 2008). A true balancing metric of raw features is the case in which the metric to match on is the full set of features for a single unit. Matching on this is a very difficult

task in high dimensions.

$$b(x) = X \tag{2.14}$$

ii. Balancing Scores. are a coarser balancing metric and is a more feasible method in higher dimensions. I refer to balancing scores as ”approximate” because these scores are of lower dimension than the original features. The features themselves form the balancing score, so matching on the score may be easier to implement than matching on all features, but will presumably reflect the same information. There are two notable balancing scores – propensity scores and prognosis scores.

Propensity Scores are the most popular balancing metric. This score is the probability of being assigned to the treatment ($T = 1$) conditional on observed features (X). Equation 2.15 (Barnow/And Others 1980; Rosenbaum/Rubin 1983b; Lechner 2001; King/Nielsen 2018).

$$e_i = Pr(T_i = 1|X_i) \tag{2.15}$$

Prognosis scores, alternatively, are the predicted outcome under the control condition. This is tantamount to modeling and matching on baseline risk, prior to treatment. Equation 2.16 (Hansen 2008; Stuart 2010; Arbogast/Ray 2009; Glynn et al. 2012; Kelcey 2013).

$$e_i = Pr(Y|X_i, T = 0) \tag{2.16}$$

Balancing scores may be estimated through logistic regression models (Rosenbaum/Rubin 1983b; D 'agostino 1998; Weitzen et al. 2004); recursive partitioning, random forests, or other tree-based models (Setoguchi et al. 2008; Lee et al. 2010; Westreich 2010; Linden/Yarnold 2017) which may include the use of bagging or boosting methods (McCaffrey et al. 2004a; Hernán/Robins 2006; Lee et al. 2010); support vector machines (Westreich 2010); and neural networks (Glynn et al. 2006; Cavuto et al. 2006; Setoguchi et al. 2008).

Distance. After we define a balancing metric, it is necessary to quantify the difference between two units. The difference – or conversely, similarity – between any two units, i and j, is called a *distance*, D_{ij}. There are different distance metrics depending on the balancing metric chosen – raw features or balancing score.

i. Distance Metrics for Raw Features. Distance calculation on the raw features is dependent on the type of data, whether its *categorical* or whether its *quadratic*. I define categorical data as that which can be represented as binary feature vectors (Cha 2007; Seung-Seok et al. 2014). There are many different metrics to quantify this distance, but a few of the most popular are Jaccard/Tanimoto and Hamming.

Tanimoto/Jaccard. In the case of binary feature vectors, Tanimoto distance and Jaccard distance are equivalent. This quantifies the ratio of the common elements to the number of all different elements (Kompan 2011; Kohonen et al. 2001; Lourenco et al. 2004; Deza/Deza 2009; Vilar et al. 2012; Goshtasby 2012).

Hamming. Measures the total mismatches of the corresponding feature categories of two units but is applicable to any ordered sets of equal length (Hamming 1950; Lourenco et al. 2004; Morris et al. 2014).

Similarly to the categorical distances, there are quadratic distances, which are continuously measured variables, such as measurements which are common in the electronic health record. Quadratic distances measure the distance between continuous feature vectors (Deza/Deza 2009; Shirkhorshidi et al. 2015). Again, there are many distance metrics for this type of data, the most common of which being Euclidean, Mahalanobis, and the Canberra Metric.

Euclidean. Measures the ”ordinary” straight-line distance between points in Euclidean space (Spiel et al. 2008; Deza/Deza 2009; Nielsen 2016).

Mahalanobis. The Euclidean distance adjusted for covariance. If there are two features that are highly correlated, then their contribution to the distances should be lower (Rubin/Thomas 2000; Rosenbaum et al. 2007; Stuart 2010; King 2011b; Baltar et al. 2014).

The Canberra Metric. Used for data scattered around an origin. This metric is sensitive to small changes when feature values are near zero (Lance/Williams 1966; Lance/Williams 1967; Kaur 2014).

For both categorical and quadratic data, the choice of distance calculation on raw features is dependent upon your research question. Relevant considerations include whether you value similarity or dissimilarity; and whether zeros hold important information. This is design decision that is likely specific to the research question, and reflects your beliefs about potentially confounding features to match on.

ii. Distance Metrics for Balancing Scores. Balancing scores are the lower dimensional, often scalar representation, of features. There are specific distance metrics to aid matching on this data, including Linear or Log Linear Distance.

Linear. The difference between the balancing scores of two units. Equation 2.17

$$D_{ij} = b(x_i) - b(x_j) \tag{2.17}$$

Log Linear. (Rubin/Thomas 1992) states that this distance supports matching on a scalar linear summary of features X, [that is a monotonic] function of the probability that a unit receives the treatment. It is noted that this is a particularly effective tool at reducing bias (Rosenbaum/Rubin 1985a; Rubin/Thomas 1992; Rubin 2001). The equation for this distance is given by 2.18, in which "logit" refers to the log-odds of the balancing score. This metric is strictly greater than zero, and is scaled, which provides constancy, a property the linear difference does not have.

$$D_{ij} = |\mathrm{logit}(b(x_i)) - logit(b(x_j))| \tag{2.18}$$

Matching Procedure. The final component of matching, are the details of the matching procedure itself. These include the (1) tolerance limit on distances, (2) the ratio of matching, and (3) whether units are replaced if matched.

1. Tolerance Limit on Distances. The tolerance limit describes how close two units must be to be considered similar. There are three domains of tolerance – (i) a distance of zero, (ii) distance within a caliper, and (iii) distance within a strata.

(i) Distance of Zero. Each unit i in the treatment group, is matched with control unit j such that the predetermined observed features of unit i, $X_i = X_j = x$ (ROSENBAUM/RUBIN 1985a; DEHEJIA/WAHBA 2002; ZHAO 2004). A Tolerance limit of zero, means the distance between two units must be exactly 0 for them to be considered a match (Equation 2.19). Assuming perfect measurement and perfect information, this would completely eliminate bias. In the case of raw features, this is an exceptionally difficult task, especially with high dimensional data, because it would involve finding an exact match for all the features (LALONDE 1986). In the case of a balancing score, this is almost impossible (IMAI et al. 2009).

$$D_{ij} = 0 \tag{2.19}$$

A version of Distance of Zero is referred to a *fine balance.* Per Rosenbaum, the term "finely balanced" is intended to suggest that the nominal variable, often with many levels, has been balanced exactly at every level, that is, with fine attention to detail (ROSENBAUM 1989; ROSENBAUM et al. 2007). This enforces balance on a variable that is difficult to balance by other methods.

(ii) Distance within a Caliper. Units are matched when balancing scores differs by less than pre-specified amount, C, called a caliper width. Equation 2.20 (Austin 2010, 2011a; Lunt 2014). This is very common when using propensity scores (Austin 2011a; Lunt 2014).

$$|D_{ij}| \leq C \tag{2.20}$$

This method is especially popular with propensity score matching in light of how difficult it is to find an exact match. But selecting the caliper width has a bias-variance trade off that can affect the quality of the match. Narrow calipers, match more similar subjects, which will decrease bias but increase variance. Wider calipers, will result in matching less similar subjects, which will increase the bias but decrease the variance (Austin 2011a).

(iii) Distance within a Strata. Sometimes called *blocking* or *stratification* (Imbens/Rubin 2015), this process matches units that are in similar ranges of the distribution of each feature or single balancing metric. Equation 2.21 (Cochran 1968; Rosenbaum/Rubin 1985a; Stuart 2010). The current convention is to use 5 strata – "quintiles" – but larger sample sizes may require more strata (Rosenbaum/Rubin 1984; D 'agostino 1998; Lunceford/Davidian 2004; Austin 2011a; Imbens/Rubin 2015).

$$b(x_i), b(x_j) \in \big[\text{Lower Bound}, \text{Upper Bound}\big] \tag{2.21}$$

2. Ratio of Matching. The second consideration in the matching procedure is the ratio of matching, which can be grouped into *pair* and *subset.* Pair matching, sometimes called *optimal pair matching* is 1:1 matching. This process selects, for each treated unit i, the single control unit, j, with the most similar balancing score or smallest distance from unit i (HANSEN/OLSEN KLOPFER 2006; STUART 2010). This is effectively, the Nearest Neighbor method. Optimal pair matching is closely aligned with the idea of matched pairs in randomized experiments, which has high power because groups are maximally similar (WACHOLDER/ WEINBERG 1982).

In the presence of lots of data, as is typical in EHRs and other observational sources, its possible to get several good control for every treatment unit (SMITH 1992; RUBIN/THOMAS 2000; STUART 2010). In this case, the ratio of matching is something I call, *Subset Matching.* These can arranged into three groups. (i) Fixed Number of Comparators, (ii) Variable Number of Comparators, and (iii) Variable Number of Treatment and Comparators.

(i) Fixed Number of Comparators. Matching one treatment unit to k control units (ROSENBAUM/RUBIN 1983b). Selecting the number of controls, k, involves a bias-variance trade-off in which increasing k will increase bias and decrease variance (AUSTIN 2010; STUART 2010).

(ii) Variable Number of Comparators. When one treatment unit is matched with a varying number of controls (MING/ROSENBAUM 2001; HANSEN/OLSEN KLOPFER 2006; STUART 2010). This is frequently used to match units within a caliper or strata. In this case, there is a bias-variance trade-off when selecting the width of the caliper. A wide caliper will increase bias and decrease variance; and a narrow caliper will decrease bias increase variance (Austin 2010, 2011b).

(iii) Variable Number of Treatment and Comparators. In which one treatment unit is matched with one or more control units, or multiple treatment units is matched with one control unit (Rosenbaum 1989, 1991; Thrusfield 2017; Hansen 2004; Hansen/Olsen Klopfer 2006; Rosenbaum 2012). Using a variable number of treatment and comparators offers efficiency gains, especially when units are scarce. A sub-method of this group, is *Cardinality Matching*, which seeks to maximize the total number of matched samples (or the cardinality), through whatever means necessary, but subject to a constraint on feature-balance (Zubizarreta et al. 2014; Keele/Zubizarreta 2014). These constraints could be (a) Weak, which requires the means on features to be balanced ("means balanced") (Zubizarreta 2012) (b) Strength-k, which is weak balance on k covariates (Hsu et al. 2015) and (c) Fine Balance, which requires the distribution of one nominal feature to be the same across arms (Rosenbaum et al. 2007).

3. Replacement. Once a control unit is selected, the matching procedure must specify whether or not the unit goes back into the pool. When units are not replaced, a single comparator unit is matched to one treatment unit. This is a greedy procedure, so the order of matching matters and the result, and it may only yield a local optimum. As a result, this may increase bias of effect estimates (Parsons 2001). Conversely, control units can be selected with replacements. Under this procedure, a single comparator unit can be matched to multiple treatment units. Under this matching procedure, the order of matching does not matter and a global optimum can be achieved (Zubizarreta 2012; Keele/Zubizarreta 2014).

Matching methods possess a number of strengths. They support flexible and robust causal modeling under selection on observables (Imai 2013). Matching separates reducing selection bias from the analysis of outcomes (Rosenbaum/Rubin 1983b; Rubin 2007), and reduces the dependence of estimates in parametric models (Ho et al. 2007). With exact matching, controlling further is unnecessary as strong ignorability is upheld (Iacus et al. 2012).

That being said, exact matches are infeasible. With increasing dimensions, matching becomes increasing difficult. Generally, matching methods do not perform as well when the features are not normally distributed or there are many features (Gu/Rosenbaum 1993; Stuart 2010). Matching is also blind to subpopulation imbalance (King/Nielsen 2018) and is inefficient, resulting in units being unmatched. If there are many unmatched units, it can result in larger bias than if the matches are inexact but more individuals remain in the analysis (Rosenbaum/Rubin 1985a; King 2011a). Matching can pose difficult a-priori assumptions, it may be hard to know if a tolerance level, caliper, or strata is reasonable (Smith/Todd 2005). Lastly, matching does not account for unobserved confounding and may introduce a bias-variance trade off.

Weighting

Weighting in the causal inference setting is similar to survey sampling re-weighting (Morgan/Todd 2008). In survey sampling, when the sample is not representative of the overall population population, units are disproportionally considered to make the sample look more like the population. Larger weights are assigned to the individuals who are under-represented in the sample, and a lower weight is assigned to those who are over-represented. Similarly, in the causal inference setting, if one treatment arm fails to look like the other, the arm may consider units differently by assigning them a *weight*. Weighting methods generalize matching methods. Similarly, all types of matching are special cases of weighting with discrete balancing weights (Imai 2013). The balancing weight dictates how much more or less we want to consider types of units, with the same weights being applied uniformly within unit type. In practice, weights are multiplied by each units' features and outcome metric. The application of the weights to cohort features results in a *pseudo population* in which the unconditional form of strong ignorability is enforced and counterfactual reasoning of the weighted outcome can take place (Czajka et al. 1992; Robins et al. 2000; Lunceford/ Davidian 2004).

It is infeasible for observational populations to have a convenient representation of types of units. How units are identified by types is being actively studied and continues to be an open research question. A review of the literature shows the following weighting methods have been used; (i) inverse probability of treatment weighting, (ii) augmented inverse probability of treatment weighting, (iii) weighting by the odds of treatment, (iv) kernel weighting/overlap weighting, and (v) coarsened exact matching

(i) Inverse Probability of Treatment Weighting. Also called inverse propensity weighting, is by far the most popular weighting method. Like matching methods, inverse probability of treatment weighting (IPTW) is a method to control for observed confounding. Despite obvious ties to propensity score matching, IPTW was developed independently (Robins 1986).

IPTWs are constructing by estimating each units' probability of having received their respective treatment, based on the observed features. Units are then weighted by the inverse of this estimated probability (Heckman et al. 1998; Dehejia/Wahba 2002; Thoemmes/Ong 2016). IPTWs are generalized in Equation 2.22 where in $T_i = 1$ is the propensity score. Like the matching, this weighting metric is very sensitive to extreme values of the propensity score, and could result in high variance or instability in these cases.

$$w_i = \frac{T_i}{P(T_i = 1|X_i)} + \frac{1 - T_i}{1 - (P(T_i = 1|X_i))} \tag{2.22}$$

To combat the instability in Equation 2.22, researchers developed a variation of this metric. It involves the baseline probability of being assigned to their treatment, which is estimated from a model with *no features*, being normalized by the probability of assigning treatment given the features. These are referred to *stabilized weights*, and tend to produce estimates with smaller variances.

Equation 2.23

$$w_i = \frac{P(T_i = 0)}{P(T_i = 1|X_i)} + \frac{1 - P(T_i = 0)}{1 - (P(T_i = 1|X_i))} \tag{2.23}$$

(ii) Augmented Inverse Propensity Weighting. Yet another variation of IPTW. This method produces consistent ATE estimates if either the outcome model or propensity model is misspecified, but the other model is correctly specified (Robins et al. 1994; Glynn/Quinn 2009). As such, the weighting technique is said to produce a doubly robust ATE model (Scharfstein et al. 1999).

(iii) Weighting by the Odds of Treatment. Though this method does not appear to be in practice, weighting by the odds of treatment is an alternative weighting technique, in which both the treatment and control groups are weighted to represent the treatment group (Hirano et al. 2003; Morgan/Todd 2008). This could be a useful technique in the case of very small treatment groups, and a larger, more heterogeneous control groups.

$$w_i = T_i + (1 - T_i)\frac{P(T_i = 1|X_i)}{1 - P(T = 1|X_i)} \tag{2.24}$$

In the weighting by the odds of treatment procedure, all units in the treatment group receive a weight of 1. And control units are weighted up to the full sample using the $1/(1P(T = 1|X_i))$, and then weighted to the treatment group using by $P(T = 1|X_i)$

(iv) Kernel Weighting & Overlap Weighting. Other methods that do not appear to be in practice. An implicit assumption of feature-balance is that the baseline risks for each group are the same. Two methods, which I found independently of one another in my literature review, both seek to equilibrate baseline risk through approximately the same means. The first is *Kernel Weighting* and the second is *Overlap Weighting*. Both of these methods seek to directly correct any differences between the non-treated potential outcome for the treatment and control groups. Put simply, these methods seek to adjust the baseline risk of the outcome to be equal across arms.

In kernel weighting, sometimes called kernel balancing, weights are then chosen on the control units such that the treated and control group have equal [means. As] a result, the expectation of the non-treatment potential outcome must also be equal for the treated and control groups after weighting (Rosenblatt 1956; Imbens 2004; Hazlett 2016).

In overlap weighting, each unit's weight is proportional to the probability of that unit being assigned to the opposite group. In theory, this method would have perfect balance of feature means. Equations 2.25 and 2.26 (Li et al. 2018).

$$w_0 \propto P(T_i = 1|X_i) \tag{2.25}$$

$$w_1 \propto 1 - P(T_i = 1|X_i) \tag{2.26}$$

(v) Coarsened Exact Matching. Though Coarsened Exact Matching (CEM) involves a fair amount of matching, this procedure also uses weighting. The idea behind CEM is to coarsen each feature, so make the recording less precise, whereupon features are matched on these coarsened values. The coarsened features are stratified, and comparators are weighted proportionally to their prevalence to the treatment group. From here, coarsened values are dropped, and the analysis is completed on uncoarsened data (Grove/Fisher 1930; King 2011b; Iacus et al. 2012; Kallus 2017).

CEM has efficiency gains over matching because unmatched units don't have to be 'thrown out.' However, CEM bounds the degree of model dependence by ex-ante user choice.

Overall, like matching methods weighting offer flexible and robust causal modeling under selection on observables (Imai/Ratkovic 2013). Unlike matching, weighting is suitable when more than two groups to compare, and the procedure is much more efficient as all units are included in the analysis (Halpern 2014). This is useful when there is no common support between arms (Crump et al. 2009). However, there is a fair amount of model dependence, which makes estimates susceptible to model misspecification. Additionally, the weights themselves are also estimated and thus have sampling variability. And like many other causal inference methods, weighting does not account for unobserved confounding (Imai/Ratkovic 2013). Most notable is that weights, particularly IPTW-based metrics, can be unstable. When the propensity score is extreme (very close to 0 or very close to 1) the weights become problematic and can leave to downstream complication with ATE estimation (Li et al. 2018).

Adjustment

Adjustment can mean many things, but in the case of causal inference, it refers to the idea of *statistical adjustment*. When treatment and control groups are not similar, we can use statistical adjustment to "control" for feature imbalance (McNamee 2005; Pourhoseingholi et al. 2012). Adjustment methods can be grouped into two categories: (i) stratification and (ii) multivariate modeling methods.

> **Stratification.** (Frangakis/Rubin 2002) states that, the basic principal stratification of [a population], P, with respect to [a set of] confounders, C, is the partition of units i = 1 ... N such that within any subset of P, all units have the same vector of confounders. In the literature, there is a distinction between adjustment for pre-treatment variables and post-treatment variables.
>
> Adjustment on pre-treatment variables accounts for potentially confounding features. Whereas adjustment on post-treatment variables – a process known as *principle stratification* – adjusts for things like adherence and loss-to-follow-up (Robins 1986; Angrist et al. 1993; Frangakis/Rubin 2002; Vanderweele 2011; Pearl 2011).
>
> Under stratification, comparator arms are split into subgroups according to the presence and absence of the confounder. When the cohort is split on all observed confounders, the ATE is taken in the subgroups for which all

confounders are absent. This method is best suited for a small study sample with a limited number of confounders. Because as the number of confounders increases, this becomes an increasingly difficult task (Imbens/Rubin 2009; Imbens et al. 2009). In this process, one could *understratify* in which you don't account for enough confounders, and produce a biased estimate; or one could *overstratify* which results is a very small number of units in each confounder-free arm, and will increase variance of effect estimates (Greenland et al. 2000; Greenland/Morgenstern 2001; Reuter 1991). For these reasons, adjustment by multivariate modeling methods is more popular.

Multivariate Modeling Methods. Multivariate modeling methods, sometimes called "response surface remodeling" is a method of reducing bias in ATE estimates from regression (Schochet 2010). states that 'we can adjust for differences between the treatment and comparators group's observable [characteristics. If] the functional form relationship between the outcome and features is specified correctly, [models can] produce unbiased estimates of [ATEs].' In practice this entails, explicitly modeling the relationship between treatment, outcome, and features (Pourhoseingholi et al. 2012; Kaplan 2018). Adjustment is, of course, useful in traditional regression settings for ATE, such as linear or logistic regression. But because these methods underlie much of the more complex machine learning methods, they can be useful with those methods too (Peters et al. 2013; Hill 2011; Zigler et al. 2012; Belloni et al. 2013; Langevin et al. 2004; Wager/Athey 2018).

In the linear regression setting. The main effect model for a relationship between an outcome, Y and an exposure, T, is given by Equation 2.27.

$$Y_i = \hat{\beta_0} + \hat{\beta_1} T_i + \epsilon_i \tag{2.27}$$

The same causal model, but adjusted for a single feature, C, is given by Equation 2.28.

$$Y_i = \hat{\beta_0} + \hat{\beta_1} T_i + \hat{\beta_2} C_i + \epsilon_i \tag{2.28}$$

It's the same as the model above, but there is an additional term for the feature. Assuming 2.28 is the true and correct specification of this causal system, including the feature, C, in the adjusted model reduces bias in the coefficient for treatment, β_1 (Neyman 1923; Splawa-Neyman et al. 1990; Rubin 1973b, 1973a, 1974, 1977, 1978; Rubin 1985; Rubin 1986; Rubin 1990; Holland/Rubin 1987; Holland 1988; Rosenbaum/Rubin 1983b; Rosenbaum/Rubin 1983a; Rosenbaum/Rubin 1984; Rosenbaum/Rubin 1985a; Rosenbaum/Rubin 1985b; Sobel 1994; Sobel 1995; Kaplan 2018). The coefficient of treatment supports ATE estimation, and provides additional insight into the direction, magnitude, and significance of the association between the treatment and the outcome (Cox 2013). When adjusting in multivariate models, one can adjust (i) by the observed features or (ii) by the propensity score.

(i) Adjustment by Observed Features. This is popular method to adjust for observed pre-treatment, potentially confounding variables. As mentioned earlier,

a confounder is an extraneous variable whose presence affects the variables being studied so that the results do not reflect the actual relationship between the variables under study (Elwood 1988; Robins/Greenland 1994; Kelsey/ Thompson 1986). Therefore, if the model is specified correctly, including potentially biasing variables as features will adjust for this distorting effect (Jager et al. 2008).

(ii) Adjustment by Propensity Scores. Alternatively, some researchers adjust on the propensity score. In theory, when the features X are sufficient to control for confounding of the effect of exposure on outcome, then adjustment for the propensity score, e_i, is also sufficient. Equation 2.28 (Rosenbaum/Rubin 1983b; Vansteelandt/Daniel 2014). Because the propensity score is scalar, this has the benefit of reducing the number of variables in the model needed to reach causal unbiasedness and may increase stability of estimates (Rosenbaum/ Rubin 1983b; D 'agostino 1998).

$$Y \perp T | e \tag{2.29}$$

Adjustment methods offer several strengths for reducing ATE estimates. They are complementary to matching methods and can be used together to produce good results (Stuart 2010). In some modelling methods, the coefficients have high interpretability. With small sample and large number of features, adjusting by propensity score can simplify adjustment (Steyer et al. 2002). One can easily check magnitude of confounding by comparing the treatment coefficient of the main effect model and adjusted model. However, adjustment methods perform poorly in the absence of common support. This scenario requires the extrapolation of ATE (Dehejia/Wahba 2002; Glazerman et al. 2003) and offers no evidence on the counterfactual state, which forces stronger reliance on our modeling (Gelman/Hill 2006). Like weighting and matching, adjustment is also susceptible to misspecification, and it requires assumption about relationships between exposure, outcome, and features. And finally, the sample size will always limit the number of degrees of freedom available for feature adjustment (Gelman/Hill 2006).

Attributable Risk Estimation

Attributable risks (AR) are probability of an exposure being a cause of an outcome; they help us interpret and communicate causal relationships between exposure and outcome (Leviton 1973; Miettinen 1974; Markush 1977). The AR may also be referred to as the population AR (Breslow/Day 1980; Boslaugh/McNutt 2008); the risk difference (Sinclair 2003), or the population etiologic fraction (Kleinbaum et al. 1982; Schlesselman/Stolley 1982).

Often in health, there is high uncertainty of exposure-outcome relationships. We leverage risks, such as the AR, to inform the risk factors for the outcomes, and to better communicate and understand the causal system. This is because ARs support the inference of whether a given outcome was *caused* by a particular exposure (Rosen 1978). Such an assessment can be made at the population level (*global* *inference*) or at the individual level (*local inference*). At the population level, the AR can be interpreted as the proportional increase in average outcome risk over a specified time interval that would be achieved when under the exposure of interest from the population while accounting for other risk factors (Rockhill et al. 1998; Greenland/Robins 1988). While at the individual level, the AR can be interpreted as the increase in outcome risk for a particular patient that would be achieved when under the exposure of interest, given that individuals other exposures. AR's local inferences could be used to inform treatment by highlighting the likely cause of an outcome for a single patient. While the global inferences may assist in identifying risk factors to be prioritized in public health policy, or the treatment of a single patient in which features are not known.

Typical estimation of AR requires knowledge of the causal graph, in which relationships between exposures, outcomes, and confounders are made explicit. But in the setting of many potential exposures – the high-throughput setting – causal graph construction may be infeasible. As an alternative, confounders to AR estimation may be controlled for through propensity-score modeling, however this may be inefficient when estimating AR for many exposures. To estimate the ARs of many exposures simultaneously without knowledge of the causal graph, the primary question is one of model specification. For this research, we explore a particular model specification for estimating attributable risks in the context of unstructured binary exposures and outcomes.

When causal graph construction is not feasible, common methods for AR estimation include (i) calculation of outcome proportion that is attributable to an exposure, using the definition of AR from (LEVIN 1953); (ii) approximation using disproportionality methods; and (iii) regression-based models.

Levin's Calculation Though many variations on AR exist, one of the earliest such formula for an exposure's AR is given by Equation 2.30, in which $P(D)$ is the prevalence of the exposure in the population and $P(D|\bar{F})$ is the conditional probability of disease among those without the exposure, F (LEVIN 1953).

$$AR = \frac{P(D) - P(D|\bar{F})}{P(D)} \tag{2.30}$$

If an exposure corresponds with an increased probability of the disease or outcome, $P(D|F) > P(D|\bar{F})$, then the AR will be bounded between 0 and 1. (CROWSON et al. 2009) This exposure is then considered a *risk factor* or a *determinant* of that outcome.

Disproportionality Analyses. DPAs are univariate methods that are often used in the analysis of spontaneous reporting data. Under a fixed time, disproportionality methods, such as risk ratios (RRs) or odds ratios (ORs), may be used to infer the unconditional AR for a cohort using Equation 2.31. (ROTHMAN 1976; CROWSON et al. 2009)

$$AR = \frac{RR - 1}{RR} \tag{2.31}$$

These statistical measures of the strength quantify the association between an exposure and an outcome by identifying exposures that co-occur with the outcome more often than expected. (Kim 2017; Schmidt/Kohlmann 2008; Zhang/Yu 1998) RRs and ORs are suitable for large counts of data, however when counts are small, estimates may become unstable. Consider, the example, the situation in which the expected number of outcomes (E) is 0.005 and the observed number of outcomes (N) is 1. In this scenario the RR would be 200, which suggests a strong association between exposure and outcome, but may reasonably have occurred by chance. (Canida 2017)

The Multi-Gamma Poisson Shrinker (MGPS) is another disproportionality method that corrects for this instability by imposing a prior that *shrinks* large RRs that stem from small counts. (Dumouchel 1999; DuMouchel/Pregibon 2001; Canida 2017). Under the MGPS, N are modeled using a Poisson likelihood (μ) and E is treated as a constant. For every exposure-outcome pair, $\langle i, j \rangle$, this model estimates $\lambda = \frac{\mu}{E}$. Both the prior and the posterior for λ are given by a mixture of two gamma distributions. An expression for this metric can be found in Equation 2.32, in which θ represents the parameteres of the two gamma distributions. The Empirical Bayes Geometric Mean (EBGM) of the posterior distribution serves as a summary statistic that may replace the RR or OR to support AR estimation. (Ihrie 2019) The MGPS-EBGM is currently the preferred method to data-mine for high association exposure-outcome relationships in Food and Drug Administration (FDA) data. (Harpaz et al. 2013;

Sakaeda et al. 2013; Ali 2011)

$$EBGM_{i,j} = e^{\mathrm{E}[\log(\lambda_{\mathrm{i,j}})|\mathrm{N_{i,j}},\theta]} \tag{2.32}$$

Both the Levin Definition and disproportionality methods, like the RR and MGPS-EBGM, provide signals that may be used to identify the risk factors that have high attributable risk. These methodologies support population-level (*global*) inferences of the exposure through AR estimation but are not suitable for risk estimation at the individual-level, in which exposures may vary.

Regression-based Methods. Regression-based models are multivariate methods that are often used in epidemiological studies where confounder-control is required. (Greenland/Drescher 1993; Coughlin et al. 1991; Kooperberg/Petitti 1991; Cox/Li 2012; Deubner et al. 1980) These methods often regress the outcome of interest by the exposures, as explanatory variables. Regression-based methods may provide individual-level (*local*) inferences of the outcome, in the form of a probability. They may also be used to support global inferences of the exposures through AR estimation, but the AR must be derived from the model coefficients. This is a straight-forward calculation in the setting of logistic regression, but may be a more complex task in other models. Even when calculated, AR estimates may be unstable and lack interpretability.

Univariate models - such the Levin Definition and DPAs – do not control for confounding, which makes AR estimates susceptible to bias. Current regression-based methods may also be insufficient, as they assume a very particular model specification that should align with causal assumptions. If specified incorrectly, regression-based models may be biased by backdoor paths between the exposure and the outcome (Pearl 1993), or subject to challenges in interpretation of regression coefficients if collinearity in the explanatory variables exists (Belsley et al. 2004). Furthermore, neither DPA nor regression-based methods may be used to all three attributable risk-related tasks: (i) global inferences of the exposure risk, (ii) local inferences of the exposure risk, and (iii) local inferences of the outcome probability.

A Brief Overview of Gaps

In reviewing these methods of causal inference, I noted the weaknesses associated with each method and identified a desiderata of features for an ideal method to generate new causal inference from observational data. The finalized desiderata includes (i) robustness to observed confounding, (ii) robustness to unobserved confounding, (iii) flexibility with heterogeneous data; (iv) efficiency, (v) scalability, (vi) model independence, and (vii) feasibility of implementation. Notable gaps in this review are an efficient, model-independent weighting method, and a method of high-throughput AR estimation.

Chapter 3

Aim 1. Determine the feasibility of observational data to replicate existing causal knowledge.

A number of techniques exist to improve and support causal estimates from observational data, but at present, there is no widely-used framework to evaluate modeling assumptions relative to experimental data. RCTs, which we accept to be the least biased source of causal knowledge, can be compared to estimates generated from observational data and, thus, provide a methodology to assess the validity of causal claims and a platform with which to evaluate inference methods. This can serve as framework for evaluating methods for causal inference. Evidence-Based Medicine (EBM) requires medical practitioners to consider empirical and experimental evidence when treating their patients. EBM encourages practitioners to seek the most reputable evidence according to a hierarchy, in which randomized controlled trials (RCTs) are regarded as the gold standard source. While RCTs can offer precise and valid insights into the efficacy and safety, the results are often criticized for their poor generalizability and may therefore be unsuitable to serve as evidence for the practice of medicine. This failure of an RCT to represent an indicated population may be the result of cohort selection for a single condition of interest that is further narrowed by a breadth of eligibility criteria. This final cohort population may stand in stark comparison to the intended target population, which could have multiple problems and comorbidities. Observational data sources, such as the electronic health record (EHR), are regarded as more representative of the target patient population; but they cannot guarantee distributional similarity on confounding variables. How and the extent to which observational data differs from experimental data is unknown.

Aim 1.1. Assess the ability to replicate causal claims from RCTs using electronic health record (EHR) data.

Background. In medicine, we often trust that biomedical research results will generalize to any given population (Ioannidis 2014; Contopoulos-Ioannidis et al. 2008; Wong/Steiner 2018). The practice of applying medical evidence to clinical care is known as Evidence Based Medicine (EBM), wherein clinicians are encouraged to consume evidence to inform the best treatment of their patients (Djulbegovic et al. 2009; Djulbegovic/Guyatt 1976; Sackett et al. 1996). Underlying the practice of EBM is the assumption that the effect shown in experimental study populations will replicate in real-word populations that any clinician sees.

The hierarchy of EBM assigns the randomized controlled trial (RCT) to the highest level of reputability (Sackett et al. 1996). A common design element of RCTs are eligibility criteria. If we consider trial patients to be a function of their features, such as demographics, laboratory measurements, and medical history; the eligibility criteria nominally define the features that all patients in a study must share. Strict RCT eligibility criteria may disallow patients with characteristics that yield more variation in the treatment effect and may be more representative of the real world.

The population that results from the application of a study's eligibility criteria may not be the same population as the study population. This may result in poor external validity – the extent to which causal relationships of a study are able to persist over variation in persons and treatment settings (Campbell/Stanley 1963). With poor external validity, study evidence, such as the average treatment effect (ATE), cannot be replicated over any population of patients (Wales 2009; Moher et al. 1996; Britton et al. 1999; Karanis et al. 2016; Stuart et al. 2015). Presumably, the eligibility criteria of a study should be sufficient to identify the precise population in which the ATE will replicate, which we call the applicable population.

When assessing replicability of a clinical study, it is important to examine whether or not the effect of the treatment varies across patient subgroups, such as those defined by age, sex, or medical history. This type of variation is known as the heterogeneity of treatment effect (HTE) (Kerbyson et al. 2014). If an experimental population differs from a real-world population among subpopulations that induce HTE, the effect estimates may not be replicated in the real-world population. In practice, it is not feasible that clinicians evaluate HTE for each patient; rather they must assume that HTE is not present across subgroups. However, when applying evidence to real-world populations, this assumption is likely unmet, as recent research has empirically proven that HTE is often found to exist (Fredriksson/Johansson 2008; Xie et al. 2012). This raises concerns for replicability of studies in highly heterogeneous, real-world populations.

When a clinician evaluates whether or not a trial's findings are applicable to a single patient, the most thorough assessment they can make employs the eligibility criteria. The eligibility criteria should nominally define the applicable population, but this assumption is only valid if the criteria capture all subpopulation heterogeneity. However, this presumption has not been tested.

We seek to assess the sufficiency of eligibility criteria in constructing a fully externally valid cohort, and the degree to which the effect estimate varies as a function of eligibility criteria. This is in contrast to research which seeks to replicate effect estimates from experimental studies with observational data, (Anglemyer et al. 2014; Hemkens et al. 2016; Franklin et al. 2017; Franklin et al. 2019; Cain et al. 2015; Hernán/Robins 2016; Bollen et al. 2015) which investigate neither how the effect estimates change with each application of eligibility criterion, nor how well the eligibility criteria identify the applicable population from the population that meets the study indication.

This study complements existing research on replicability of effect estimates from experimental studies with observational data (Anglemyer et al. 2014; Hemkens et al. 2016; Franklin et al. 2019; Franklin et al. 2017; Franklin/Schneeweiss 2017; Cain et al. 2015; Hernán/Robins 2016; Bollen et al. 2015; Open Science Collaboration 2015). We investigate how similar the applicable real-world population is to the clinical trial population and assess HTE in the excluded population through an analysis of eligibility criteria. This research bridges the knowledge gap around the sufficiency of eligibility criteria in constructing what should be a fully externally-valid cohort.

Research Questions. *Does the construction of observational cohorts according to RCT eligibility criteria encourage the causal effect estimate to converge with that reported in the trial?*

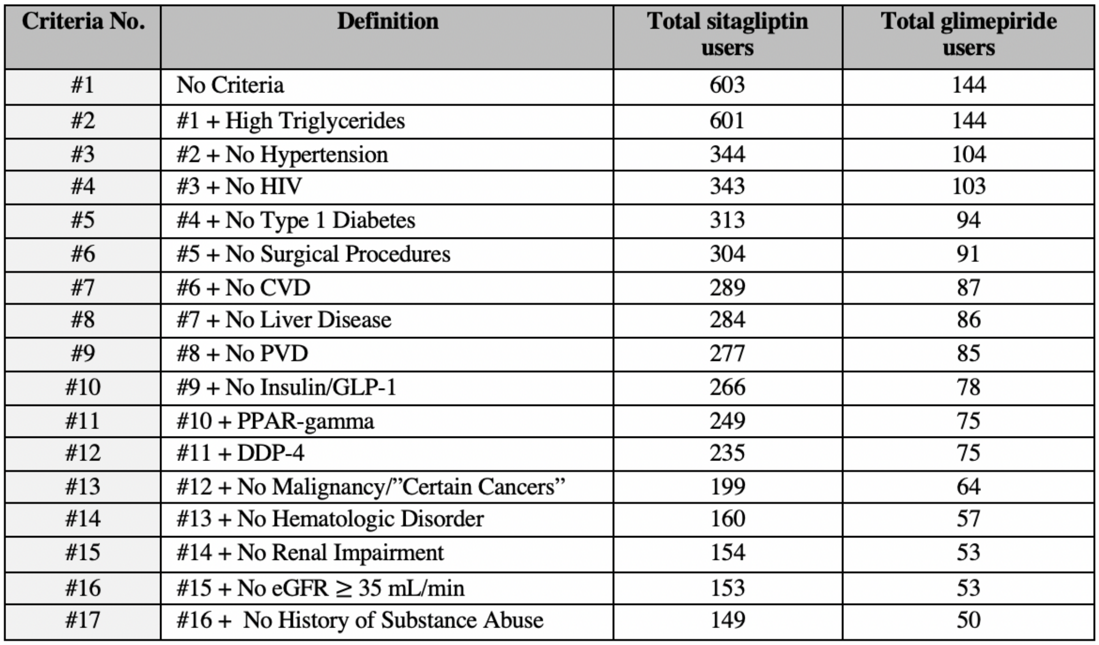
Methods. We hypothesized that HTE is present in real-world populations and could be demonstrated through an analysis of eligibility criteria. Specifically, we evaluated whether each addition of an inclusion criterion would increase effect estimate similarity between the observational cohort and the published RCT estimates. To address this, we applied RCT eligibility criteria to EHR data, used state-of-the-art methods to control for confounding, and evaluated local causal estimates. To curate the observational cohorts, we leveraged the OHDSI ATLAS cohort-creation tool (Hripcsak et al. 2015).

We first constructed a baseline study population, that was defined by the indication of the study drug and any age and gender restrictions put forth by the trial. To be eligible for the baseline cohort, patients must have the target indication, have no inpatient or outpatient use of either drug under comparison, and initiate treatment in an outpatient setting. This is a restrictive requirement but makes treatment naivety very likely in the context of observational data. Inclusion and exclusion criteria extracted for each trial were identified as the union of criteria from (i) publications, (ii) protocols, and (iii) clinicaltrails.gov, and were operationalized using the OHDSI CDM. Operationalization was done as faithfully as possible and intended to mirror the interpretation of clinicians. To ensure the fidelity of this process, operationalization was supervised by a clinician. All criteria were operationalized except for soft criteria, which we define as criteria that cannot be reasonably put into a computable form, e.g. current participation in another study in the prior 12 weeks or expected survival of two years.

Inclusion and exclusion criteria extracted from published RCT protocols and literature were incrementally added to an observational cohort that was constructed according to the baseline study population of the comparison RCT. With each incremental criterion, effect estimates for endpoints associated with each RCT were calculated and compared to the published RCT results. Unadjusted effect estimates and effect estimates with propensity score matched cohorts were both calculated. Adjusted effects were estimated using the OHDSI CohortMethod package (Schuemie et al. 2016). Propensity scores were estimated using a logistic regression, regularized with a Laplace prior, and fit using a large collection of conditions, procedures, medications, and measurements excluding the outcomes of interest. We matched units in a 1:1 fashion on the propensity score, using a caliper of 0.2. The data were analyzed in a main-effect outcome model with no covariates. For this research, the outcome model used was a logistic regression. This model was selected to produce the same effect estimates (odds ratios) that were reported in the literature.

This method was applied to the RCT, *Efficacy and Tolerability of sitagliptin Compared with glimepiride in Elderly Patients with Type 2 Diabetes Mellitus and*

Inadequate Glycemic Control: A Randomized, Double-Blind, Non-Inferiority Trial, which evaluated the occurrence of hypoglycemia associated with sitagliptin compared to glimepiride in elderly patients (65-80 years of age) with Type 2 Diabetes Mellitus (T2DM) and inadequate glycemic control (Hartley et al. 2015). This trial will be queried for the endpoints of (i) composite serious adverse-events (SAEs); (ii) hypoglycemia; (iii) HbA1c $< 7.0\%$; and (iv) HbA1c $< 6.5\%$. The eligibility criteria for this RCT are detailed in Table 3.1. Details on how the interventions and outcome were defined and codified can be found in the Appendix for Aim 1.2.



Criteria No.	Definition	Total sitagliptin users	Total glimepiride users
#1	No Criteria	603	144
#2	#1 + High Triglycerides	601	144
#3	#2 + No Hypertension	344	104
#4	#3 + No HIV	343	103
#5	#4 + No Type 1 Diabetes	313	94
#6	#5 + No Surgical Procedures	304	91
#7	#6 + No CVD	289	87
#8	#7 + No Liver Disease	284	86
#9	#8 + No PVD	277	85
#10	#9 + No Insulin/GLP-1	266	78
#11	#10 + PPAR-gamma	249	75
#12	#11 + DDP-4	235	75
#13	#12 + No Malignancy/”Certain Cancers”	199	64
#14	#13 + No Hematologic Disorder	160	57
#15	#14 + No Renal Impairment	154	53
#16	#15 + No eGFR ≥ 35 mL/min	153	53
#17	#16 + No History of Substance Abuse	149	50

Figure 3.1: Sequential eligibility criteria and resultant counts in each cohort, associated with the Hartley, et al RCT.

Data. The observational clinical data for our cohorts will be obtained from the NewYork-Presbyterian Hospital (NYPH) clinical data warehouse (CDW). The CDW contains observational clinical data for 5.37 million individual subjects from 1986-2017.

Patients generated during outpatient, inpatient, and emergency room visits. Data will be formatted according to the OHDSI common data model (CDM) to support downstream interoperability of our models within the OHDSI community and to promote their adoption by OHDSI collaborators. Data modalities used to train our models will include clinical labs, medications, procedures, and diagnosis codes. Cohort size under each incremental criterion can be found in Table 3.1.

Evaluation. The goal of this Aim is to best attempt to replicate RCT trial results with observational data. In the event that the effect estimates are discordant, we assume that the underlying patient populations are dissimilar or that there exists a heterogeneity of treatment effect for some eligibility criteria.

Results. The results of this Aim suggest the presence of HTE in the excluded population. The change from the Indication Only cohort (subject to no eligibility criteria) to the Indication+Eligibility Criteria cohort (subject to all eligibility criteria), highlights the lack of HTE for some variables but the presence of HTE for others.

The presence of HTE in the excluded population is demonstrated in If there is very little variation in the observational effect estimate when subject to different eligibility criteria, this may be indicative of a lack of HTE (homogeneity of treatment effect). This potential lack of HTE in the excluded population is seen in the outcomes of composite endpoint of SAEs, HbA1c $< 6.5\%$, and HbA1c $< 7.0\%$. For these outcomes, both the unadjusted and matched analysis of the data under increasing criteria resulted in minimal changes to the effect estimate. However, the outcome of hypoglycemia

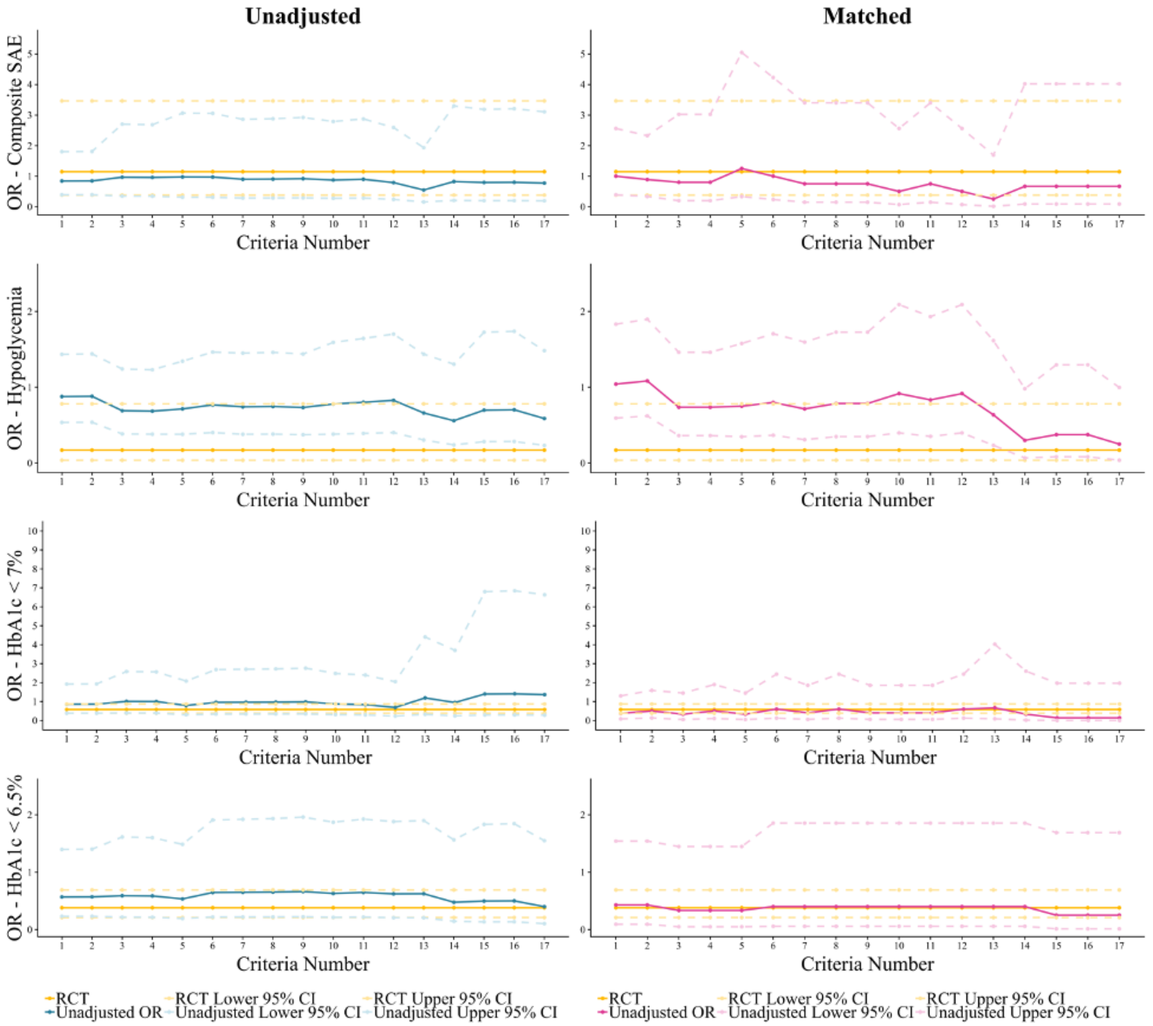


Figure 3.2: Odds ratio for (i) composite serious adverse-events (SAEs); (ii) hypoglycemia; (iii) HbA1c < 7.0%; and (iv) HbA1c < 6.5% under increasing eligibility criteria.

shows greater changes to the effect estimate when patients that do not meet the eligibility criteria are removed. If removal of these patients greatly changed the effect estimate, than this may indicate that more heterogeneous outcomes were also excluded.

Though the effect estimate for all outcomes demonstrates a trend towards the reported RCT effect estimate, the results presented here are underpowered to evaluate the significance of this trend (Figure 1). The odds ratio (OR) for the composite endpoint of SAEs showed a non-significant harmful effect of sitagliptin versus glimepiride on a composite endpoint of serious adverse events ($OR_{RCT} = 1.15$ [0.38, 3.46]). Our observational results, though also non-significant, indicate that contrary results to the trial – that glimepiride had comparatively higher risk for this outcome ($OR_{Unadjusted} = 0.77$ [0.20, 3.12] and $OR_{Matched} = 0.67$ [0.09, 4.02]). The effect of sitagliptin versus glimepiride on hypoglycemia was found to be significant in the trial ($OR_{RCT} = 0.17$ [0.04, 0.78]), which indicates a protective effect of sitagliptin for this outcome. When all criteria were applied, the observational ORs similarly found a protective effect of sitagliptin. Of the three observational effect estimates for hypoglycemia, adjustment by matching most closely resembled the trial ($OR_{Matched} = 0.25$ [0.04, 1.00] and $OR_{Unadjusted} = 0.59$ [0.23, 1.48]). The RCT found a significant protective effect of sitagliptin as compared to glimepiride for both efficacy endpoints, HbA1c $<$ 6.5% ($OR_{RCT} = 0.38$ [0.21, 0.69]) and HbA1c $<$ 7.0% ($OR_{RCT} = 0.58$ [0.38, 0.87]). Application of the eligibility criteria appear to be less impactful on the effect estimates of efficacy – resulting in smaller changes to the OR – than on effect estimates of adverse effects. For the endpoint of HbA1c $<$ 7.0%, $OR_{Matched} = 0.14$ [0.000, 1.97] and $OR_{Unadjusted} = 1.36$ [0.28, 6.64]) which indicates that patients taking sitagliptin were more likely to achieve these efficacy endpoints. For the endpoint of HbA1c $<$ 6.5%, all observational estimates were similarly protective to the trial ($OR_{Unadjusted} = 0.34$ [0.10, 1.55] and $OR_{Matched} = 0.25$ [0.01, 1.69]).

Discussion. The variation of effect estimate as a function of the eligibility criteria suggests that there exists a HTE among the real-world subpopulations. Based on these results, careful consideration beyond eligibility criteria is necessary to determine whether results of a given RCT are an appropriate source of evidence when considering the care of a given patient.

As a consequence of this empirical demonstration, this research suggests that eligibility criteria are not sufficient for identifying the applicable real-world population in which experimental treatment effects will replicate. The observational effect estimates under all eligibility criteria were all marked with varying degrees residual bias that could not be accounted for by our outlined procedure. Because we carefully identified real-world populations according to all of the trial eligibility criteria, we assume that our procedure identifies the correct patients from our observational data source. However, given the failure of replication that was seen, it is possible that the identification of patients by eligibility criteria fails to create a cohort in which the correct patients occur in the correct proportions. The RCT effect estimates that were presented in this Aim, are an average of individual treatment effects – trial participants may have a treatment effect below or above this average number (Gabler et al. 2009). Such variability is also known as the HTE (Duan/Wang 2012). For perfect replication with observational data, the distribution of treatment effect must be the same as that in the trial. Therefore, the inability of perfect replication may due attributable to the HTE.

Through our replicability efforts, we were also able to articulate a framework of external validity. As noted earlier, external validity refers to the extent to which the trial results can be applied outside of the experimental setting.15 The results of Study 1 demonstrate that we were never able to achieve perfect replication, despite high-fidelity application of criteria and adjustment for potential confounders through stratification and matching. As a consequence of this failure, we may assert that the results of the trials in Study 1 are not externally valid for our population. This process can be applied to any study of an intervention and comparator, and may provide valuable insight into the generalizability of causal knowledge.

Limitations. This research does have limitations. The translation of clinical trial eligibility criteria to operationalized and computable queries may be prone to subjectivity. Though we sought to represent the criteria as unbiasedly as possible and consulted with clinicians to ensure accuracy, there is inherent ambiguity in the criteria themselves, which make perfect RCT representation impossible. Furthermore, information regarding the eligibility criteria may be found within the clinical note, which was not used when constructing the cohorts in this research. The effect estimates show in Aim 1.1 were learned from a single sites data. The results we show, may therefore, not be generalizable to either other observational data sources or the RCT. Finally, as noted before, it is likely that the results presented in Aim 1.1 are underpowered. The underpowering of this study likely stems from low patient counts in this study; as we required that patients both be inpatient, treatment naïve, and meet strict eligibility criteria. As a consequence, the results are underpowered for the detection of causal relationships. As such, the empirical results presented must be interpreted with caution.Lastly, and most notably, experimental data and EHR data are fundamentally different, which makes comparison between these two sources difficult. Though differences to experimental data may be inherent, the EHR houses the information that is available to clinicians at the time when treatment decisions are made. Furthermore, it is a valuable resource for identifying the applicable patients to support the practice of EBM. We believe that discrepancies between experimental data and EHR data are necessary to study so that we may develop methodologies to ensure appropriate applicability at the point of care.

Aim 1.2. Examine potential sources of residual bias in effect estimates from observational data sources.

Background. Generalizability closes the gap between biomedical research and clinical practice (Wong/Steiner 2018). When research is translated into the healthcare setting, the application of biomedical evidence to clinical care is known as Evidence Based Medicine (EBM). Since its inception in the 1990s, EBM has become the standard of operation for many clinicians (Djulbegovic/Guyatt 2017a, 2017b; Djulbegovic et al. 2009; Sackett et al. 1996). EBM encourages clinicians to seek the most reputable evidence for any patient, according to a hierarchy of study quality in which randomized controlled trials (RCTs) are the best single study design (Sackett et al. 1996). RCTs are most often used to unbiasedly assess the effect of an intervention, such as a drug or procedure, on an outcome.

Though EBM may be employed successfully for many different clinical decisions, challenges remain. Underlying EBM's success is the assumption that the effect shown in RCTs will replicate in real-world populations (Ioannidis 2014; Contopoulos-Ioannidis et al. 2008). However, research has shown that factors beyond the intervention itself, such as age, sex, or medical history, may modify the measured effect, a phenomenon known as heterogeneity of treatment effect (HTE) (Kerbyson et al. 2014). If the RCT population differs from the real-world population based on factors that induce HTE, RCT results will not be replicated in real-world application. Realistically, clinicians cannot evaluate HTE on a case-by-case basis and must assume that HTE is not a significant factor. However, when applying evidence from RCTs, this assumption is likely unmet. Research has shown that HTE is often found to exist (Fredriksson/Johansson 2008; Xie et al. 2012). This raises concerns for reproducibility of studies in the presence of additional heterogeneity in real-world populations.

The RCT is well-regarded for many reasons, but randomization is the most important. Randomization ensures the highest possible internal validity, which speaks to whether the true effect is biased by systematic error (Campbell/Stanley 1963; Burns et al. 2011). The notion of internal validity does not speak to how well the causal relationship will generalize, only how unbiased it is for the study population. The patients for which the effect estimate is internally valid are nominally defined by eligibility criteria. These criteria both stipulate the characteristics that all study patients must share and nominally identify the real-world population for which the effect is internally valid. When operationalized, the eligibility criteria are represented as inclusion and exclusion criteria; (Campbell/Stanley 1963; Hyman 1982; Anderson-Cook 2005), and with every addition of a criterion to a study population, a different sub-population is identified with increasingly controlled conditions (Velasco 2010). If HTE exists, then application of eligibility criteria to a population may identify a subpopulation of patients for which there is a more homogeneous effect of the intervention.

RCTs often employ very restrictive eligibility criteria and are often cited as poorly representative of the real-world, as many subpopulations may be excluded. This may result in poor external validity. External validity refers to the extent to which the treatment effect estimate applies those outside of the study with potentially different patient and treatment setting characteristics (Campbell/Stanley 1963). External validity always poses a concern, except in the circumstance in which HTE is known to be absent.

With poor external validity, replication of the study effect can be challenging (Wales 2009; Moher et al. 1996; Britton et al. 1999; Karanis et al. 2016; Stuart et al. 2015). Replication of trial evidence with real-world data, ideally, requires that the right persons, in the right treatment setting, exist in the right proportions. In the context of treating a population that differs significantly from the clinical trial population, it can be unclear how appropriate the evidence is for this new population. Presumably, the eligibility criteria of a study should be sufficient to identify the population in which the effect will replicate, which we call the applicable population.

To address this knowledge gap, we leverage observational data to assesses if RCT populations and real-world populations after application of eligibility criteria differ. If the populations differ, the evidence may not apply due to HTE. If HTE exists in observational populations, it may impede the replication of RCT effect estimates.

These methods will contribute (i) a means to determine if the eligibility criteria are adequate for identifying the applicable population; (ii) a framework for evaluating the external validity of studies; and (iii) highlight tensions between assumptions of EBM practice and qualities of reputable evidence.

This research may encourage clinicians to reconsider the assumptions made when practicing EBM, and whether these assumptions are valid. Furthermore, the empirical evidence put forth by this study highlights the limitations of the current system of clinical knowledge generation. The current system sacrifices external validity in favor of internal validity, through the selection of the experimental population. Such a decision impedes the ability of experimental evidence to translate to the general population, resulting in non-optimal or damaging clinical care. This problem motivates the use of study populations that are more representative of the real-world and is only truly optimized when study populations and the populations targeted for treatment are one in the same. Such an analysis is called real-world evidence (RWE) generation, in which clinical knowledge is learned from the analysis of routinely-collected, real-world data (Sherman et al. 2016). The results of this research identify the need for RWE in clinical medicine and underscore how RWE may improve the practice of EBM.

Research Questions. *Why do observational causal effect estimates fail to converge with that reported in the RCT?*

Methods. We hypothesized that significant baseline characteristic differences exist between clinical trial populations and observational cohorts that meet all eligibility criteria. Such differences could be the source of poor external validity in the presence of HTE. The presence of differences could be confirmed by comparing empirical distributions of features between the RCT data and real-world observational data. However, patient-level RCT data is rarely released, so such as assessment is infeasible for most published RCTs. The best available proxy is to compare the real-world observational cohort to the summary of baseline characteristics of RCTs, as commonly presented in Table 1 of RCT publications. We will refer to these summary statistics as baseline characteristics. The baseline characteristics summarize the baseline demographic and clinical characteristics for each arm of the study (Schulz et al. 2010). The intent of publishing this table is to describe the clinical trial population in detail and report the similarity of arms in the RCT post-randomization. This data can also be used to evaluate external validity, and by association, replicability (Furler et al. 2012). To examine how potential differences between experimental and observational cohorts may contribute to poor replicability, we compared RCT baseline characteristics with the same metrics from observational EHR data.

Data. Observational clinical data was obtained from the Columbia University Irving Medical Center (CUIMC) clinical data warehouse (CDW). Data elements evaluated in this study include laboratory measurements, diagnosis codes, and medications. This database is comprised predominantly of emergency and inpatient visits with a smaller number of outpatient visits at the hospital's teaching clinics. The data used for this research was formatted according to the Observational Health Data Sciences and Informatics' common data model (CDM) to support downstream interoperability within the OHDSI community and to support replication and extension by OHDSI collaborators.

Cohort Creation. Corresponding to each RCT, observational cohorts were curated from EHR data according to two approaches. The first approach curated based on only the indication of the drug (Indication Only), e.g. diabetes or heart failure. This cohort represents the most basic assessment that clinicians can make when considering a treatment for a patient, per EBM. The second approach curated based on both the indication of the drug and all published eligibility criteria (Indication+Eligibility Criteria). This cohort represents the most thorough assessment that clinicians can make under EBM. Both the Indication Only and Indication+Eligibility Criteria cohorts were constructed using OHDSI's ATLAS tool. ATLAS is an analytics platform used to support the design and execution of observational analyses. Part of this platform includes the ability to create cohort definitions. Cohort definitions identify a set of patients that satisfy one or more criteria for a duration of time. The Indication Only and Indication+Eligibility Criteria cohorts were defined using this tool. The indication and eligibility criteria that were extracted from published RCT documentation were operationalized using the OMOP CDM and served as criteria for cohort definitions. This was a rigorously done procedure, in which medical doctors were consulted to ensure the accuracy of the operationalization and faithfulness to the original criteria. To operationalize the criteria, we created concept sets, which enumerate both the medical concepts that should be included in the definition of our criteria. and excludes the concepts that should not be included. This procedure often employed the hierarchical relationships that exist in the OMOP CDM ontology, wherein all descendants of a single concept could be selected as part of a concept set and selectively removed, if needed. This procedure is outlined in Figure 3.3.

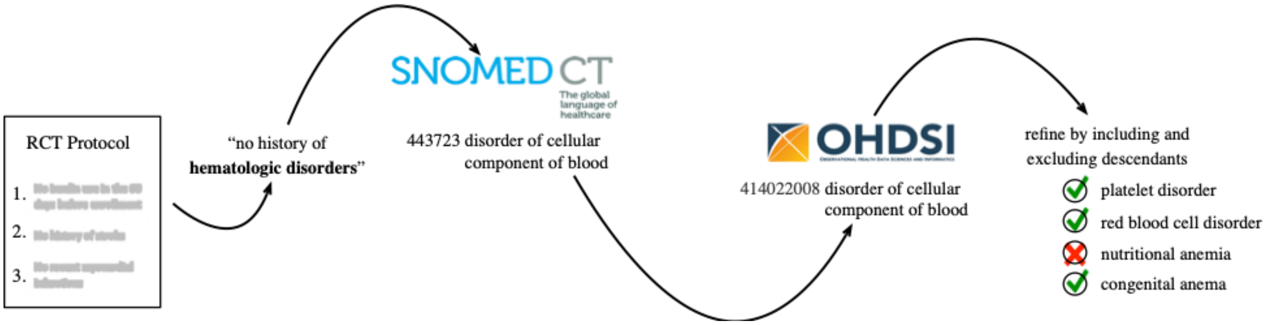
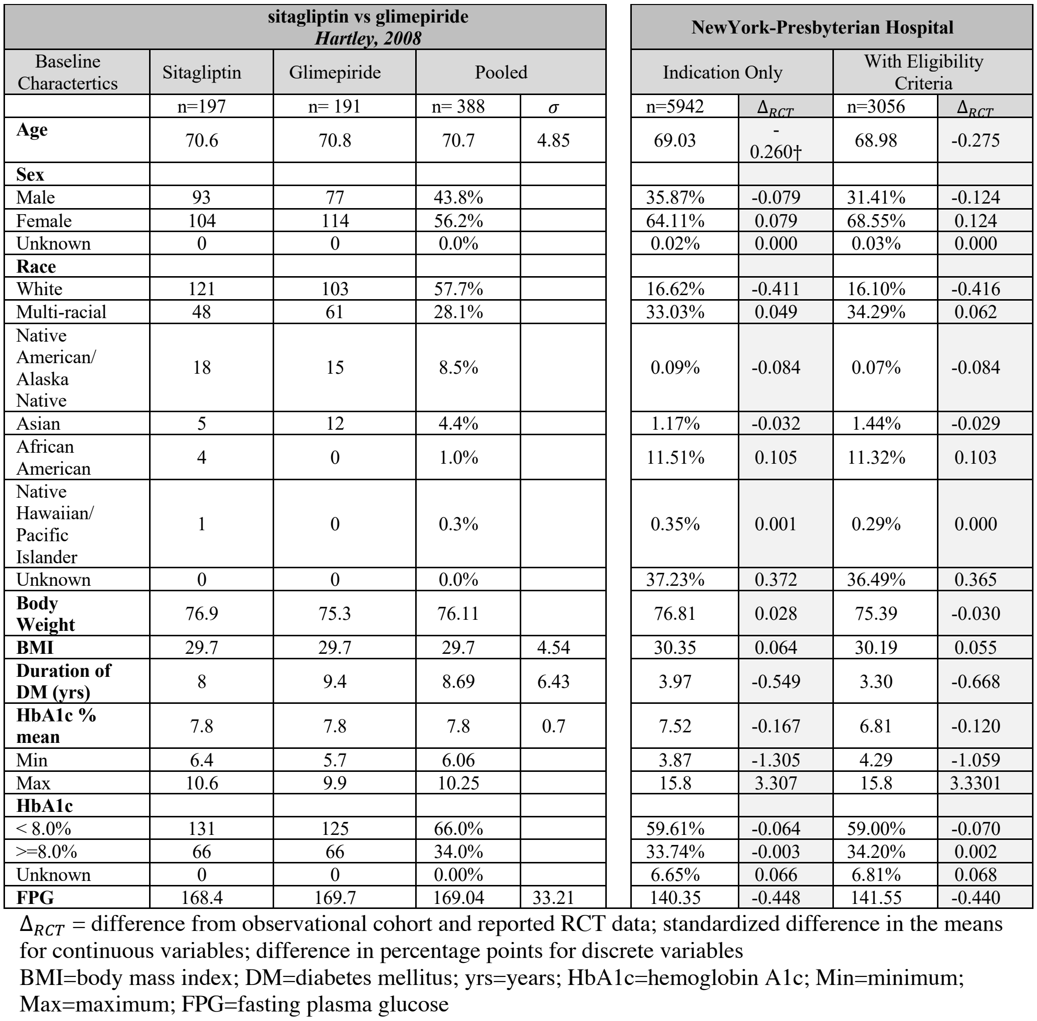

RCT Protocol
1.
2.
3.
"no history of hematologic disorders"
SNOMED CT
The global language of healthcare
443723 disorder of cellular component of blood
OHDSI
414022008 disorder of cellular component of blood
refine by including and excluding descendants
✓ platelet disorder
✓ red blood cell disorder
✗ nutritional anemia
✓ congenital anema

Figure 3.3: Pipeline to operationalize eligibility criteria using OHDSI tools. The process begins by identifying the resources (e.g., an RCT protocol) that detail the eligibility criteria of a trial. Each criterion is then extracted and mapped to codified concepts in a controlled vocabulary. The concept is then mapped to the OHDSI common data model, which aggregates the same concepts from different vocabularies, into a single standardized concept. This concept is then refined to best define the eligibility criterion.

Trial Selection. For this research, we purposefully picked landmark clinical trials, which are highly influential studies that are noted to change the practice of medicine. We began with a list of landmark trials, and after application of criteria that are outlined below, we decided on a small number. Our primary focus was landmark trials, but to increase the diversity of studies and to demonstrate applicability outside of efficacy trials, we evaluated a safety trial that met our criteria as well. When selecting candidate trials for this research, there were practical considerations that informed our choice of trials (Bartlett et al. 2019). The RCT must have an active intervention and comparator drug, as we would be unable to sufficiently codify a cohort exposed to a placebo. Additionally, the intervention cannot be a new investigatory drug, as it would not exist in our EHR. The eligibility criteria for the RCT must be published and accessible; and most of the eligibility criteria must be hard criteria that are easily operationalized into concept codes (e.g., “age of at least 55 years”). While most trials have inescapable soft criteria that are not easily operationalized (e.g., “no contraindications” or “no current participation in another clinical trial”), it is important that our chosen trials have few of these. Consider, for example, the soft criteria “expected survival of at least two years” which embodies a judgment call by a healthcare practitioner that cannot reasonably replicated with data. Finally, we sought trials that detailed a patient population that exists within the CUIMC EHR. This would ensure that a sufficient number of patients remain in our cohorts after application of the eligibility criteria. Because we are interested in comparing the RCT Table 1 metrics with the same metrics from our observational cohort, it is important that our observational data contain as many patients as possible, as greater number of patients will increase confidence that our reported data is truly representative of the CUIMC population. To that end, we investigated four trials (1) the RENAAL Trial, which compared the effect of losartan and placebo on diabetic nephropathy (Ishii 1972); (2) the ACCOMPLISH Trial (Jamerson et al. 2008), which compared benazepril-amlodipine to benazepril and hydrochlorothiazide on CV-related mortality, (3) the PROVE-IT Trial (Cannon et al. 2004), which compared atorvastatin and pravastatin in patients with a history of acute coronary syndrome (ACS); and (4) the sitagliptin and glimepiride trial (Hartley et al. 2015), which compared sitagliptin and glimepiride in elderly, diabetic patients. RENAAL, ACCOMPLISH, and PROVE-IT are Landmark RCTs with efficacy endpoints, and the sitagliptin versus glimepiride trial is a smaller trial with a safety endpoint. Details on how the Indication Only and Indication+Eligibility Criteria cohorts were created can be found in the Appendices for Aim 1.1 and 1.2

Evaluation. For each RCT under study, we calculated the pooled baseline characteristics using the metrics reported for both the intervention and comparator arms. Discrete data was summed across both arms and is presented as a percent. Continuous data was taken as the average of each arm's reported metrics, weighted by the proportion of patients in that arm. The Indication Only and Indication+Eligibility Criteria cohorts were queried to obtain metrics that corresponds to the RCT baseline characteristics. To explore the differences that exist between the observational patient cohorts and the RCT patient cohort, we calculated (i) the standardized difference in the means for continuous variables and (ii) percentage point differences between discrete variables (Δ_{RCT}). If Δ_{RCT} evaluates to zero, this indicates that the observational cohort does not differ from the trial cohort. If Δ_{RCT} does not equal zero, this indicates that observational and trial cohorts differ, with greater magnitudes corresponding to greater discrepancies between the cohorts.

Results. The results of this Study are presented in Tables 3.4, 3.5, 3.6, and 3.7 and Figure 3.8.

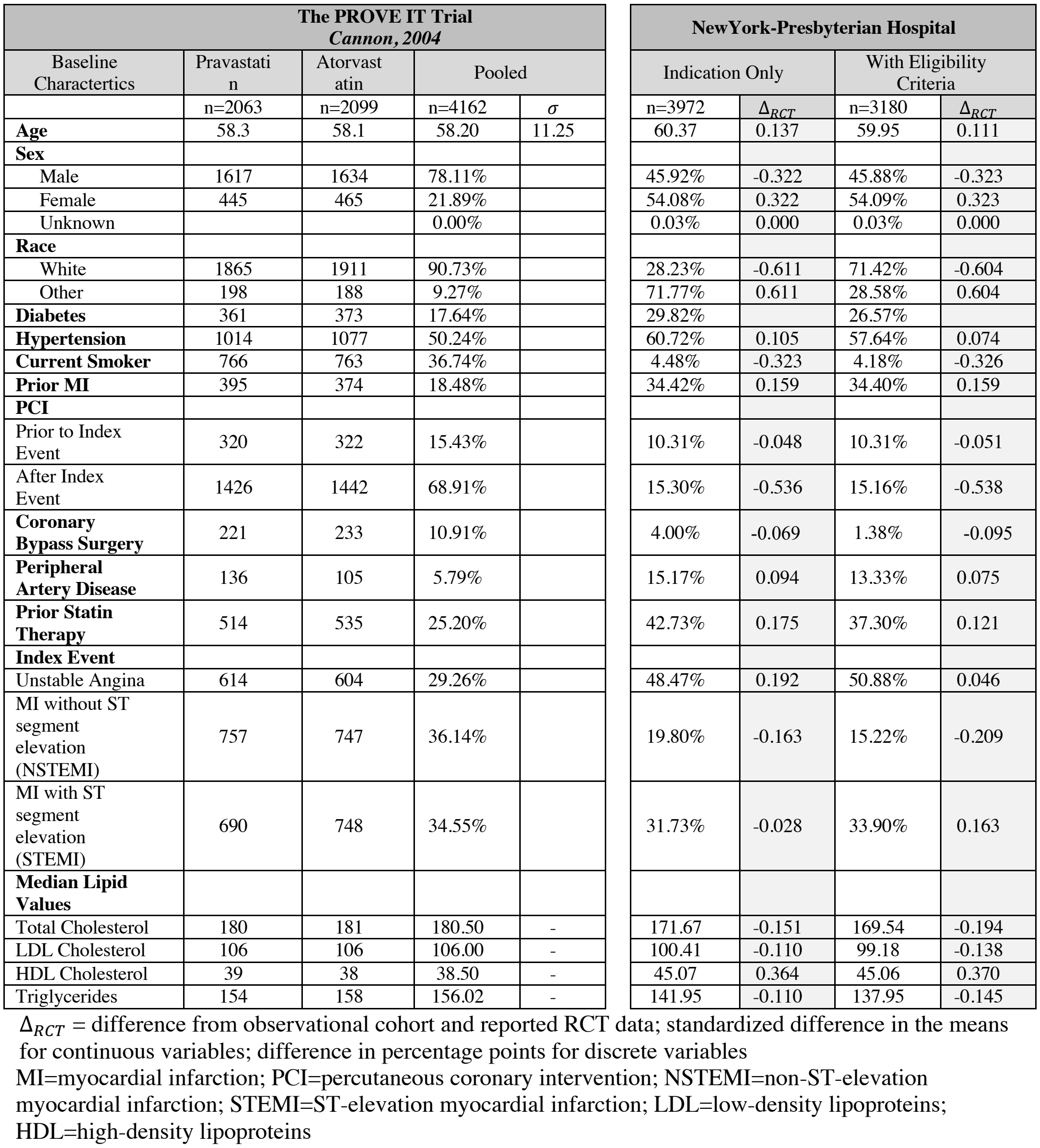


sitagliptin vs glimepiride *Hartley, 2008*					NewYork-Presbyterian Hospital			
Baseline Charactertics	Sitagliptin	Glimepiride	Pooled		Indication Only		With Eligibility Criteria	
	n=197	n= 191	n= 388	σ	n=5942	Δ_{RCT}	n=3056	Δ_{RCT}
Age	70.6	70.8	70.7	4.85	69.03	- 0.260†	68.98	-0.275
Sex								
Male	93	77	43.8%		35.87%	-0.079	31.41%	-0.124
Female	104	114	56.2%		64.11%	0.079	68.55%	0.124
Unknown	0	0	0.0%		0.02%	0.000	0.03%	0.000
Race								
White	121	103	57.7%		16.62%	-0.411	16.10%	-0.416
Multi-racial	48	61	28.1%		33.03%	0.049	34.29%	0.062
Native American/ Alaska Native	18	15	8.5%		0.09%	-0.084	0.07%	-0.084
Asian	5	12	4.4%		1.17%	-0.032	1.44%	-0.029
African American	4	0	1.0%		11.51%	0.105	11.32%	0.103
Native Hawaiian/ Pacific Islander	1	0	0.3%		0.35%	0.001	0.29%	0.000
Unknown	0	0	0.0%		37.23%	0.372	36.49%	0.365
Body Weight	76.9	75.3	76.11		76.81	0.028	75.39	-0.030
BMI	29.7	29.7	29.7	4.54	30.35	0.064	30.19	0.055
Duration of DM (yrs)	8	9.4	8.69	6.43	3.97	-0.549	3.30	-0.668
HbA1c % mean	7.8	7.8	7.8	0.7	7.52	-0.167	6.81	-0.120
Min	6.4	5.7	6.06		3.87	-1.305	4.29	-1.059
Max	10.6	9.9	10.25		15.8	3.307	15.8	3.3301
HbA1c								
< 8.0%	131	125	66.0%		59.61%	-0.064	59.00%	-0.070
>=8.0%	66	66	34.0%		33.74%	-0.003	34.20%	0.002
Unknown	0	0	0.00%		6.65%	0.066	6.81%	0.068
FPG	168.4	169.7	169.04	33.21	140.35	-0.448	141.55	-0.440

Δ_{RCT} = difference from observational cohort and reported RCT data; standardized difference in the means for continuous variables; difference in percentage points for discrete variables
BMI=body mass index; DM=diabetes mellitus; yrs=years; HbA1c=hemoglobin A1c; Min=minimum; Max=maximum; FPG=fasting plasma glucose

Figure 3.4: Results for sitagliptin versus glimepiride trial.

Sitagliptin vs Glimepiride: The sitagliptin versus glimepiride trial in elderly patients with Type 2 Diabetes Mellitus is given in Table 1. Application of eligibility criteria to the Indication Only cohort identified the Indication+Eligibility Criteria cohort that was more similar to the RCT with regard to BMI, Fasting Plasma Glucose, and HbA1c % (mean); and less similar to the RCT with regard to Age, Years Since Diabetes Diagnosis, Gender, HbA1c>8%, Race/Ethnicity, and Weight. Indication+Eligibility



The PROVE IT Trial *Cannon, 2004*					NewYork-Presbyterian Hospital			
Baseline Charactertics	Pravastatin	Atorvastatin	Pooled		Indication Only		With Eligibility Criteria	
	n=2063	n=2099	n=4162	σ	n=3972	Δ_{RCT}	n=3180	Δ_{RCT}
Age	58.3	58.1	58.20	11.25	60.37	0.137	59.95	0.111
Sex								
Male	1617	1634	78.11%		45.92%	-0.322	45.88%	-0.323
Female	445	465	21.89%		54.08%	0.322	54.09%	0.323
Unknown			0.00%		0.03%	0.000	0.03%	0.000
Race								
White	1865	1911	90.73%		28.23%	-0.611	71.42%	-0.604
Other	198	188	9.27%		71.77%	0.611	28.58%	0.604
Diabetes	361	373	17.64%		29.82%		26.57%	
Hypertension	1014	1077	50.24%		60.72%	0.105	57.64%	0.074
Current Smoker	766	763	36.74%		4.48%	-0.323	4.18%	-0.326
Prior MI	395	374	18.48%		34.42%	0.159	34.40%	0.159
PCI								
Prior to Index Event	320	322	15.43%		10.31%	-0.048	10.31%	-0.051
After Index Event	1426	1442	68.91%		15.30%	-0.536	15.16%	-0.538
Coronary Bypass Surgery	221	233	10.91%		4.00%	-0.069	1.38%	-0.095
Peripheral Artery Disease	136	105	5.79%		15.17%	0.094	13.33%	0.075
Prior Statin Therapy	514	535	25.20%		42.73%	0.175	37.30%	0.121
Index Event								
Unstable Angina	614	604	29.26%		48.47%	0.192	50.88%	0.046
MI without ST segment elevation (NSTEMI)	757	747	36.14%		19.80%	-0.163	15.22%	-0.209
MI with ST segment elevation (STEMI)	690	748	34.55%		31.73%	-0.028	33.90%	0.163
Median Lipid Values								
Total Cholesterol	180	181	180.50	-	171.67	-0.151	169.54	-0.194
LDL Cholesterol	106	106	106.00	-	100.41	-0.110	99.18	-0.138
HDL Cholesterol	39	38	38.50	-	45.07	0.364	45.06	0.370
Triglycerides	154	158	156.02	-	141.95	-0.110	137.95	-0.145

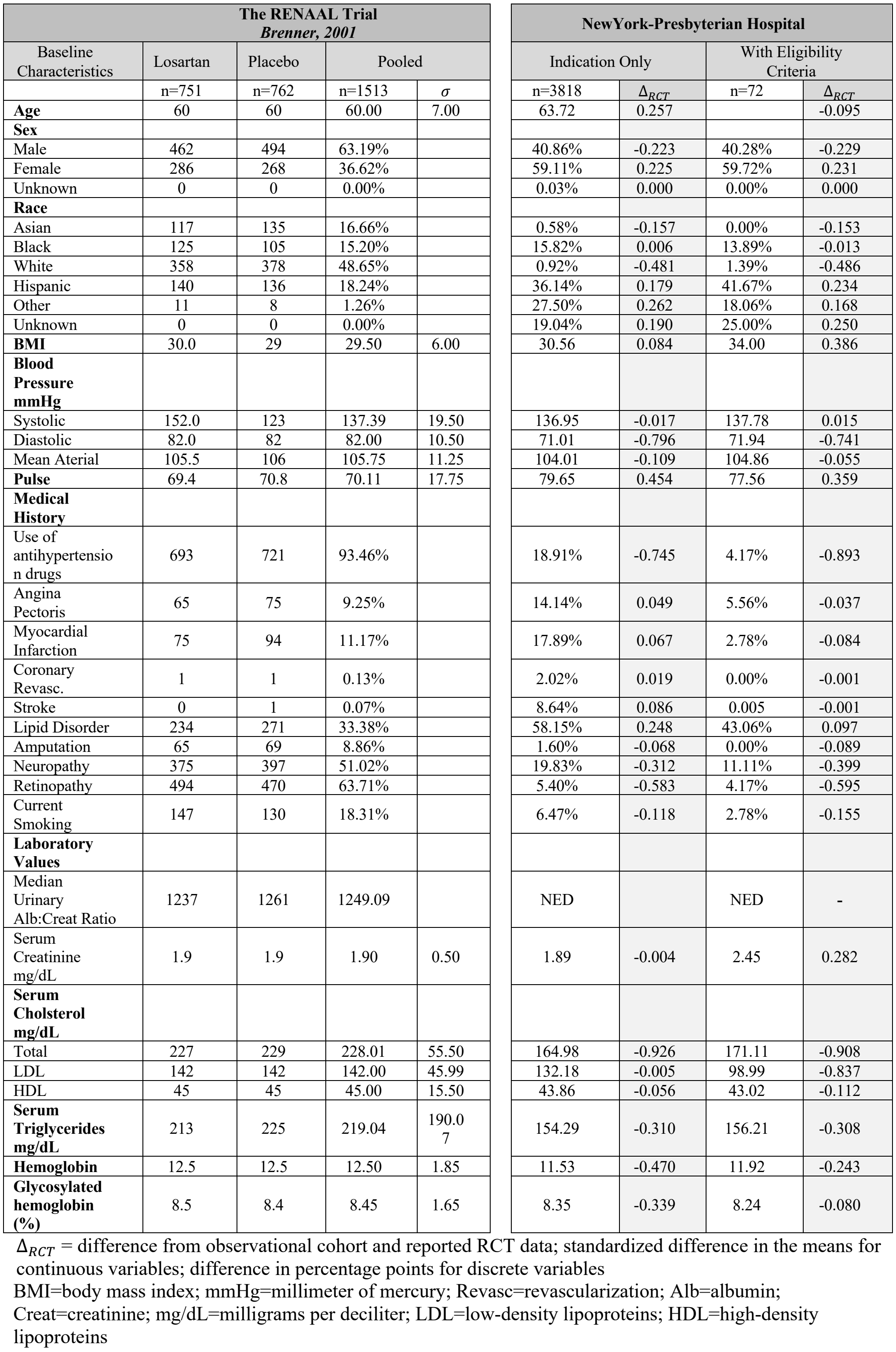
Δ_{RCT} = difference from observational cohort and reported RCT data; standardized difference in the means for continuous variables; difference in percentage points for discrete variables

MI=myocardial infarction; PCI=percutaneous coronary intervention; NSTEMI=non-ST-elevation myocardial infarction; STEMI=ST-elevation myocardial infarction; LDL=low-density lipoproteins; HDL=high-density lipoproteins

Figure 3.5: Results for atorvastatin vs pravastatin trial (PROVE-IT).

Criteria patients did not significantly differ from the trial in regards to BMI, Weight, and HbA1c % (mean), all other baseline characteristics metrics did significantly differ.

These results highlight that the indicated real-world population and the real-world population that meets the stringent eligibility criteria have generally less progressed

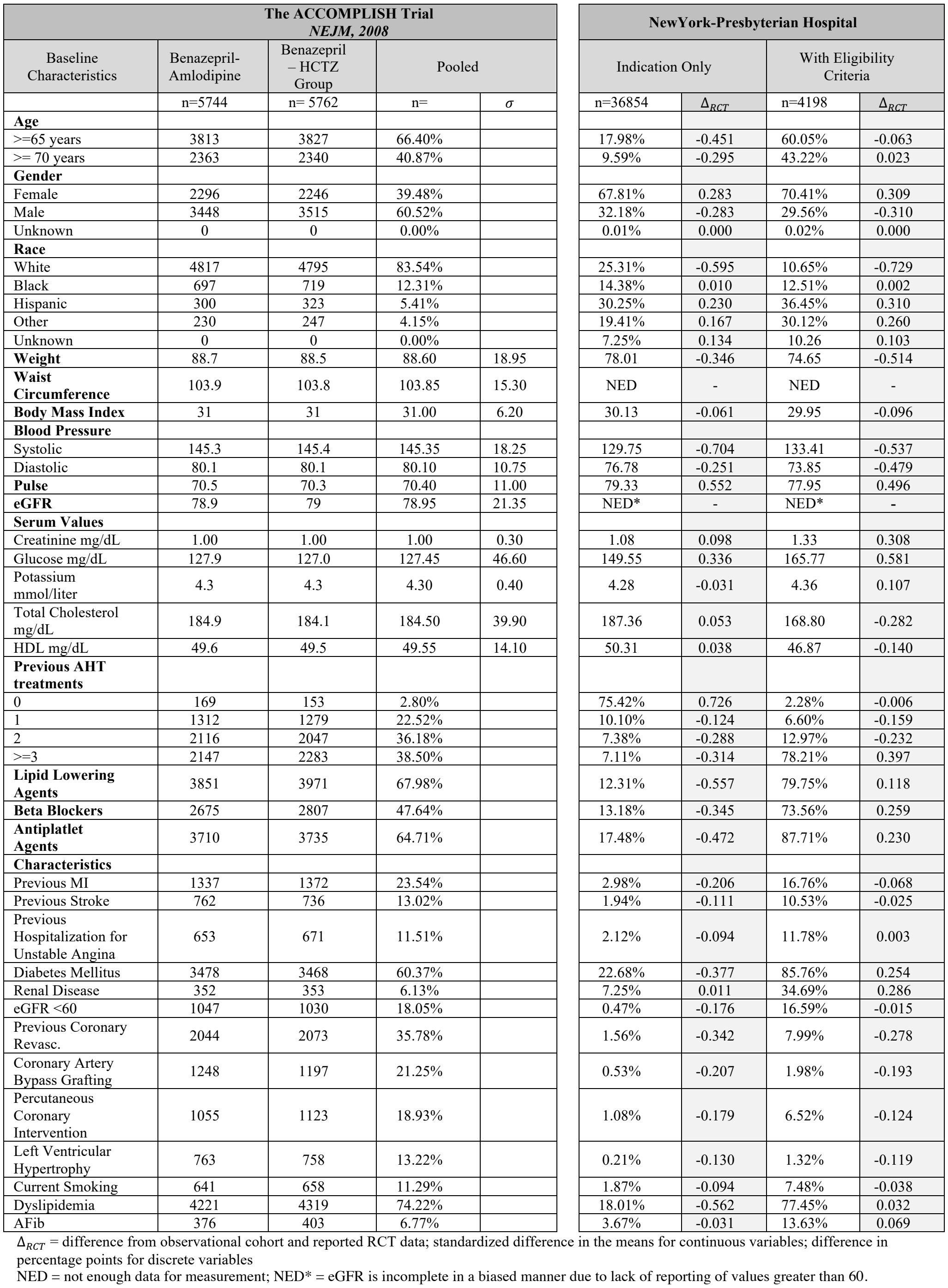


The RENAAL Trial ***Brenner, 2001***					**NewYork-Presbyterian Hospital**			
Baseline Characteristics	Losartan	Placebo	Pooled		Indication Only		With Eligibility Criteria	
	n=751	n=762	n=1513	σ	n=3818	Δ_{RCT}	n=72	Δ_{RCT}
Age	60	60	60.00	7.00	63.72	0.257		-0.095
Sex								
Male	462	494	63.19%		40.86%	-0.223	40.28%	-0.229
Female	286	268	36.62%		59.11%	0.225	59.72%	0.231
Unknown	0	0	0.00%		0.03%	0.000	0.00%	0.000
Race								
Asian	117	135	16.66%		0.58%	-0.157	0.00%	-0.153
Black	125	105	15.20%		15.82%	0.006	13.89%	-0.013
White	358	378	48.65%		0.92%	-0.481	1.39%	-0.486
Hispanic	140	136	18.24%		36.14%	0.179	41.67%	0.234
Other	11	8	1.26%		27.50%	0.262	18.06%	0.168
Unknown	0	0	0.00%		19.04%	0.190	25.00%	0.250
BMI	30.0	29	29.50	6.00	30.56	0.084	34.00	0.386
Blood Pressure mmHg								
Systolic	152.0	123	137.39	19.50	136.95	-0.017	137.78	0.015
Diastolic	82.0	82	82.00	10.50	71.01	-0.796	71.94	-0.741
Mean Aterial	105.5	106	105.75	11.25	104.01	-0.109	104.86	-0.055
Pulse	69.4	70.8	70.11	17.75	79.65	0.454	77.56	0.359
Medical History								
Use of antihypertension drugs	693	721	93.46%		18.91%	-0.745	4.17%	-0.893
Angina Pectoris	65	75	9.25%		14.14%	0.049	5.56%	-0.037
Myocardial Infarction	75	94	11.17%		17.89%	0.067	2.78%	-0.084
Coronary Revasc.	1	1	0.13%		2.02%	0.019	0.00%	-0.001
Stroke	0	1	0.07%		8.64%	0.086	0.005	-0.001
Lipid Disorder	234	271	33.38%		58.15%	0.248	43.06%	0.097
Amputation	65	69	8.86%		1.60%	-0.068	0.00%	-0.089
Neuropathy	375	397	51.02%		19.83%	-0.312	11.11%	-0.399
Retinopathy	494	470	63.71%		5.40%	-0.583	4.17%	-0.595
Current Smoking	147	130	18.31%		6.47%	-0.118	2.78%	-0.155
Laboratory Values								
Median Urinary Alb:Creat Ratio	1237	1261	1249.09		NED		NED	-
Serum Creatinine mg/dL	1.9	1.9	1.90	0.50	1.89	-0.004	2.45	0.282
Serum Cholsterol mg/dL								
Total	227	229	228.01	55.50	164.98	-0.926	171.11	-0.908
LDL	142	142	142.00	45.99	132.18	-0.005	98.99	-0.837
HDL	45	45	45.00	15.50	43.86	-0.056	43.02	-0.112
Serum Triglycerides mg/dL	213	225	219.04	190.07	154.29	-0.310	156.21	-0.308
Hemoglobin	12.5	12.5	12.50	1.85	11.53	-0.470	11.92	-0.243
Glycosylated hemoglobin (%)	8.5	8.4	8.45	1.65	8.35	-0.339	8.24	-0.080

Δ_{RCT} = difference from observational cohort and reported RCT data; standardized difference in the means for continuous variables; difference in percentage points for discrete variables

BMI=body mass index; mmHg=millimeter of mercury; Revasc=revascularization; Alb=albumin; Creat=creatinine; mg/dL=milligrams per deciliter; LDL=low-density lipoproteins; HDL=high-density lipoproteins

Figure 3.6: Results for losartan vs placebo trial (RENAAL).



The ACCOMPLISH Trial *NEJM, 2008*					NewYork-Presbyterian Hospital			
Baseline Characteristics	Benazepril-Amlodipine	Benazepril – HCTZ Group	Pooled		Indication Only		With Eligibility Criteria	
	n=5744	n= 5762	n=	σ	n=36854	Δ_{RCT}	n=4198	Δ_{RCT}
Age								
>=65 years	3813	3827	66.40%		17.98%	-0.451	60.05%	-0.063
>= 70 years	2363	2340	40.87%		9.59%	-0.295	43.22%	0.023
Gender								
Female	2296	2246	39.48%		67.81%	0.283	70.41%	0.309
Male	3448	3515	60.52%		32.18%	-0.283	29.56%	-0.310
Unknown	0	0	0.00%		0.01%	0.000	0.02%	0.000
Race								
White	4817	4795	83.54%		25.31%	-0.595	10.65%	-0.729
Black	697	719	12.31%		14.38%	0.010	12.51%	0.002
Hispanic	300	323	5.41%		30.25%	0.230	36.45%	0.310
Other	230	247	4.15%		19.41%	0.167	30.12%	0.260
Unknown	0	0	0.00%		7.25%	0.134	10.26	0.103
Weight	88.7	88.5	88.60	18.95	78.01	-0.346	74.65	-0.514
Waist Circumference	103.9	103.8	103.85	15.30	NED	-	NED	-
Body Mass Index	31	31	31.00	6.20	30.13	-0.061	29.95	-0.096
Blood Pressure								
Systolic	145.3	145.4	145.35	18.25	129.75	-0.704	133.41	-0.537
Diastolic	80.1	80.1	80.10	10.75	76.78	-0.251	73.85	-0.479
Pulse	70.5	70.3	70.40	11.00	79.33	0.552	77.95	0.496
eGFR	78.9	79	78.95	21.35	NED*	-	NED*	-
Serum Values								
Creatinine mg/dL	1.00	1.00	1.00	0.30	1.08	0.098	1.33	0.308
Glucose mg/dL	127.9	127.0	127.45	46.60	149.55	0.336	165.77	0.581
Potassium mmol/liter	4.3	4.3	4.30	0.40	4.28	-0.031	4.36	0.107
Total Cholesterol mg/dL	184.9	184.1	184.50	39.90	187.36	0.053	168.80	-0.282
HDL mg/dL	49.6	49.5	49.55	14.10	50.31	0.038	46.87	-0.140
Previous AHT treatments								
0	169	153	2.80%		75.42%	0.726	2.28%	-0.006
1	1312	1279	22.52%		10.10%	-0.124	6.60%	-0.159
2	2116	2047	36.18%		7.38%	-0.288	12.97%	-0.232
>=3	2147	2283	38.50%		7.11%	-0.314	78.21%	0.397
Lipid Lowering Agents	3851	3971	67.98%		12.31%	-0.557	79.75%	0.118
Beta Blockers	2675	2807	47.64%		13.18%	-0.345	73.56%	0.259
Antiplatlet Agents	3710	3735	64.71%		17.48%	-0.472	87.71%	0.230
Characteristics								
Previous MI	1337	1372	23.54%		2.98%	-0.206	16.76%	-0.068
Previous Stroke	762	736	13.02%		1.94%	-0.111	10.53%	-0.025
Previous Hospitalization for Unstable Angina	653	671	11.51%		2.12%	-0.094	11.78%	0.003
Diabetes Mellitus	3478	3468	60.37%		22.68%	-0.377	85.76%	0.254
Renal Disease	352	353	6.13%		7.25%	0.011	34.69%	0.286
eGFR <60	1047	1030	18.05%		0.47%	-0.176	16.59%	-0.015
Previous Coronary Revasc.	2044	2073	35.78%		1.56%	-0.342	7.99%	-0.278
Coronary Artery Bypass Grafting	1248	1197	21.25%		0.53%	-0.207	1.98%	-0.193
Percutaneous Coronary Intervention	1055	1123	18.93%		1.08%	-0.179	6.52%	-0.124
Left Ventricular Hypertrophy	763	758	13.22%		0.21%	-0.130	1.32%	-0.119
Current Smoking	641	658	11.29%		1.87%	-0.094	7.48%	-0.038
Dyslipidemia	4221	4319	74.22%		18.01%	-0.562	77.45%	0.032
AFib	376	403	6.77%		3.67%	-0.031	13.63%	0.069

Δ_{RCT} = difference from observational cohort and reported RCT data; standardized difference in the means for continuous variables; difference in percentage points for discrete variables

NED = not enough data for measurement; NED* = eGFR is incomplete in a biased manner due to lack of reporting of values greater than 60.

Figure 3.7: Results for benazepril-amlodipine vs benazepril and hydrochlorothiazide (HCTZ) trial (ACCOMPLISH).

diabetes than those patients in the trial. This is exemplified by (i) Years Since Diabetes Diagnosis, which is 3.97 for the Indication Only cohort and 3.30 in the Indication+Eligibility Criteria cohort, but is 8.69 in the trial (p=0.007) and (ii) Fasting Plasma Glucose, which is 140.35 in the Indication Only cohort and 141.55 in the Indication+Eligibility Criteria cohort, but is 169.04 in the trial (p=0.007). With regard to these two baseline characteristics metrics, the application of the eligibility criteria to the Indication Only cohort identified a subset of patients with a Fasting Plasma Glucose that was more similar to the trial and a Years Since Diabetes Diagnosis that was less similar to the trial.

PROVE-IT: The atorvastatin versus pravastatin trial in patients with a history of ACS (PROVE-IT Trial) is given in Table 2. Application of eligibility criteria to the Indication Only cohort identified the Indication+Eligibility Criteria cohort that was more similar to the RCT with regard to Age, Race/Ethnicity, Diabetes, Hypertension, Prior MI, Peripheral Artery Disease, and Prior Statin Therapy, and less similar to the RCT with regard to Sex, Current Smoker, Percutaneous Coronary Intervention, Index Event, and Median Lipid Values. Indication+Eligibility Criteria patients differed significantly from the trial in regards to all baseline characteristics.

The results for this trial show that patients that meet either the Indication or the Indication subject to all criteria, have less severe cardiovascular lipid measurements than patients in the trial. This is demonstrated in the Median Lipid Values, where in Total Cholesterol, LDL, HDL, and Triglycerides are 171.67, 100.41, 45.07 and 141.95, respectively, in the Indication Only cohort and 169.55, 99.19, 45.07, and 138.00, respectively, in the Indication+Eligibility Criteria. This is compared to the 180.50, 106.00, 38.50, and 156.02, respectively, that is reported in the trial.

RENAAL: The losartan versus placebo trial in patients with diabetic nephropathy (RENAAL Trial) is given in Table 3. Application of eligibility criteria to the Indication Only cohort identified the Indication+Eligibility Criteria cohort that was more similar to the RCT with regard to Age, Pulse, Angina Pectoris, Coronary Revascularization, Stroke, Lipid Disorder, Total Cholesterol, Serum Triglycerides, Hemoglobin, and Glycosylated Hemoglobin, and less similar to the RCT with regard to Sex, Race/Ethnicity, Blood Pressure measurements, Use of Antihypertensive Drugs, Myocardial Infarction, Amputation, Neuropathy, Retinopathy, Current Smoking, Laboratory Values, LDL and HDL. Indication+Eligibility Criteria patients significantly differ from the trial in regards to Angina Pectoris, Stroke, Amputation, Lipid Disorder, Glycosylated Hemoglobin % all other baseline characteristics metrics significantly differ. Significance of Median Urinary Alb:Creatinine Ratio measurements could not be assessed due to insufficient reporting in the EHR.

Similar to the trial results previously mentioned, patients enrolled in the RCT demonstrate hallmarks of advanced disease. A greater proportion of trial patients had a medical history of amputation (8.86%), neuropathy (51.02%), and retinopathy (63.71%), than compared to either the Indication Only cohort (1.60%, 19.83%, 5.40%, respectively) or the Indication+Eligibility Criteria cohort (0.00%, 11.11%, 4.17%).

ACCOMPLISH: The benazepril-amlopidine versus benazepril-hydocholorothiazide trial in patients with systolic hypertension (ACCOMPLISH Trial) is given in Table 4. Application of eligibility criteria to the Indication Only cohort identified the Indication+Eligibility Criteria cohort that was more similar to the RCT with regard to Age, Potassium, Lipid Lowering Agents, Beta Blockers, Antiplatelet Agents; History of MI, Stroke, Hospitalization for Unstable Angina, Diabetes Mellitus, eGFR<60, Coronary Revascularization, CABG, PCI, Left Ventricular Hypertrophy, Current Smoking, Dyslipidemia, and AFib, and less similar to the RCT with regard to Sex, Race/Ethnicity, Weight, Blood Pressure Measurements, Pulse, Creatinine, Glucose, Total Cholesterol, HDL, and History of Renal Disease. Indication+Eligibility Criteria patients significantly differ from the trial in regards to all baseline characteristics, except for history of Previous Hospitalization for Unstable Angina. Significance of Waist Circumference and eGFR could not be assessed due to data availability and insufficient reporting in the EHR.

The results of the four trials are summarized in Figure 1. In this Figure, each quadrant of the plot corresponds to a trial. For each trial, the Δ_{RCT} for baseline characteristics are plotted for Indication Only vs RCT and Indication+Eligibility Criteria vs RCT. The minimum and maximum HbA1c measurements for the NCT01189890 trial were excluded in this plot due to biologically implausible values that were likely transcription errors.

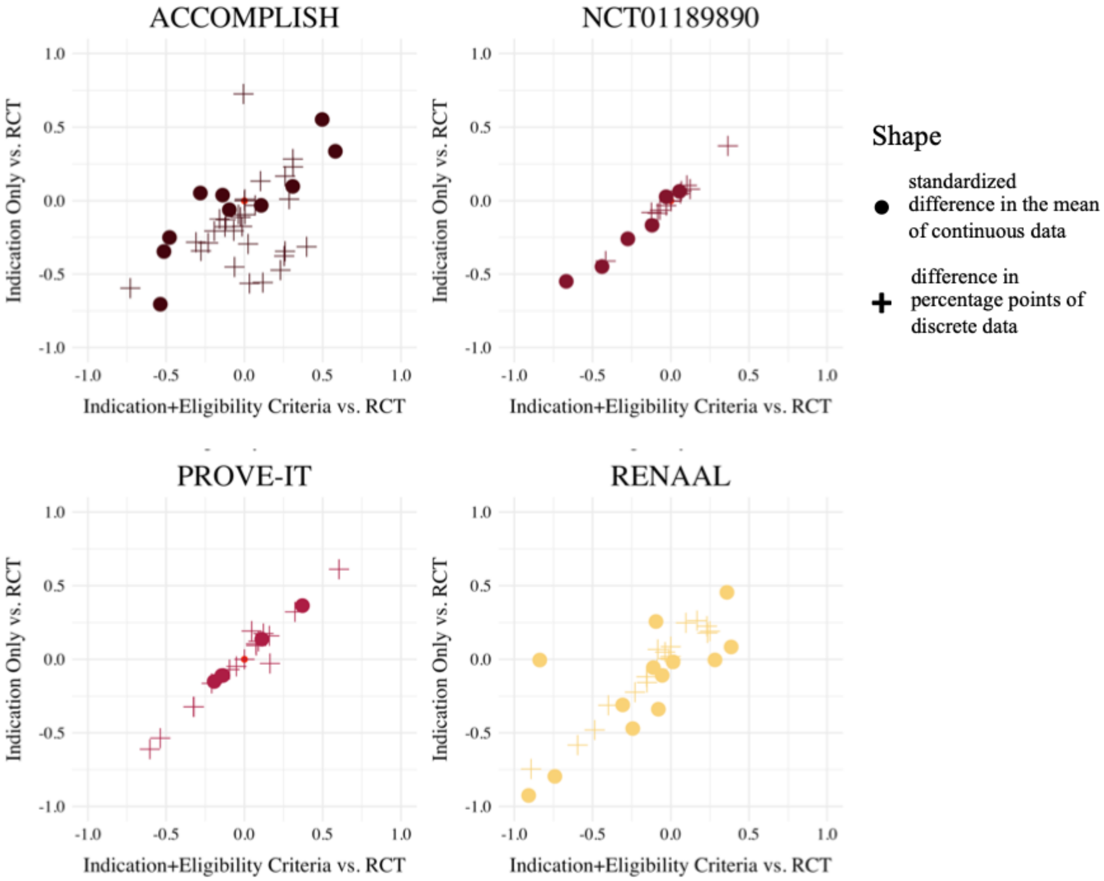


Figure 3.8: Summary of Δ_{RCT} for baseline characteristics of Indication Only vs RCT and Δ_{RCT} Indication+Eligibilty Criteria vs RCT

Discussion. Based on the results of the research presented, the eligibility criteria, that nominally should be sufficient for effect replication, may not actually be sufficient if HTE exists. If HTE exists and the differences we observed in our cohort are common, factors beyond eligibility criteria may be necessary to identify applicable patients. This finding has significant implications on how we create and apply biomedical evidence.

The expectation of EBM is that the population of patients that a single clinician sees, is an applicable population, and will mirror the population in the RCT in all ways including the distribution of the treatments effect. This assumption does not take into account variation undocumented factors that affect HTE. If factors that induce HTE are not accounted for in the eligibility criteria but exist, a clinician cannot reasonably assume that the treatment effect will be seen in his treated patient population. The discrepancies between experimental and real-world populations that are presented here may be due to a number of sources, including overly restrictive eligibility criteria, insufficient documentation of eligibility criteria, or the self-selection of trial participants. When seeking to rectify this gap and improve generalizability of RCT findings, these issues may be addressed by the relaxation of trial eligibility criteria, a thorough and accurate description of eligibility criteria (perhaps recorded in a codified manner), or the active recruitment of a representative experimental population. Regardless of the source of this discrepancy, until addressed, careful consideration beyond who is eligible for the trial is necessary to determine whether results of a given RCT are an appropriate source of evidence when considering the care of a given patient.

Limitations. This research does have limitations. Most importantly, the trials presented in this research were selected according to a set of criteria that enabled their analysis using the tools described. These criteria included an active intervention and comparator, published eligibility criteria, and ease of operationalization of concepts. The trials that were investigated as part of this research represent common indications. It is possible that the results presented here are specific to trials of common conditions and may not be representative of rare condition trials.

As in Aim 1.1, (i) the operationalization of clinical trial eligibility criteria into a computable format may be prone to subjectivity; (ii) and the EHR data may be inherently different than the experimental data. Though we attempted to translate the eligibility criteria as faithfully as possible into the observational setting, errors may have been made. In the presence of many eligibility criteria, small discrepancies between the intended meaning of a criterion and the operatonalized criterion may be compounded. Additionally, when subjecting an observational cohort to many criteria, the resultant cohort may become very small, leading to a lack of power for the detection of relevant differences. In our evaluation of the external validity of trials, we compare aggregate metrics rather than a full distribution of features, which would be preferable. This comparison is the best we can do with the data that is available to us. However, such a comparison may fail to capture meaningful differences between the trial and real-world populations, as distributions with greatly differing functional forms may still have similar means.

Chapter 4

Aim 2. Develop a method to identify natural experiments within observational data.

Causal inference often relies on the counterfactual framework, which requires that treatment assignment is independent of the outcome, known as strong ignorability. Approaches to enforcing strong ignorability in causal analyses of observational data include weighting and matching methods. Effect estimates, such as the *average treatment effect* (ATE), are then estimated as expectations under the reweighted or matched distribution, P. The choice of P is important and can impact the interpretation of the effect estimate and the variance of effect estimates. In this work, instead of specifying P, we learn a distribution that simultaneously maximizes coverage and minimizes variance of ATE estimates. In order to learn this distribution, this Aim proposes a generative adversarial network (GAN)-based model called the Counterfactual χ-GAN (cGAN), which also learns feature-balancing weights and supports unbiased causal estimation in the absence of unobserved confounding. Our model minimizes the Pearson χ^2-divergence, which we show simultaneously maximizes coverage and minimizes the variance of importance sampling estimates. To our knowledge, this is the first such application of the Pearson χ^2-divergence. We demonstrate the effectiveness of cGAN in achieving feature balance relative to established weighting methods in simulation and with real-world medical data.

Aim 2.1. Implement a generative adversarial network, Counterfactual χ-GAN, to learn balancing weights.

Background. Causal assessment often relies on the framework of *counterfactual inference*. In this framework, each unit, i, has a *potential outcome* given that they received a treatment and a potential outcome given that they received a control – $Y_{1,i}$ and $Y_{0,i}$, respectively. This framework seeks to contrast the outcome, Y for an individual under these two hypothetical states as shown in Eq. 4.1 (Rubin 1974).

$$ITE = Y_1 - Y_0 \tag{4.1}$$

The effect of the treatment on the outcome can then summarized by calculating population-level effect estimates, such as the average treatment effect (ATE), which is defined as the expected difference in outcomes (Eq. 4.2).

$$ATE = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] \tag{4.2}$$

Estimating this requires access to the outcome for the state in which units were not assigned (i.e., $\mathbb{E}[Y_0|T = 1]$ and $\mathbb{E}[Y_1|T = 0]$). In practice, however, these true counterfactuals are never observed as a single population (or individual) cannot simultaneously be both treated and untreated. This is known as the 'fundamental problem of causal inference.' Therefore, approximations that employ more than one population are used as a proxy for these unobserved states (HOLLAND 1986). These approximations seek to construct populations such that the observed ATE, $A\hat{T}E$, equals the true ATE that would arise from a counterfactual population. In other words, we seek an $A\hat{T}E$ that is *unbiased*.

$$A\hat{T}E = \mathbb{E}[Y_1|T=1] - \mathbb{E}[Y_0|T=0] \tag{4.3}$$

A decomposition of the ATE, demonstrates that a sufficient condition for unbiased $A\hat{T}E$ estimation is that $\mathbb{E}[Y_1|T=1] = \mathbb{E}[Y_1|T=0]$ and $\mathbb{E}[Y_0|T=0] = \mathbb{E}[Y_0|T=1]$ (KEMPTHORNE 1955). Within the counterfactual framework, this equality is central to the assumption of *strong ignorability* (Eq. 4.4) (ROSENBAUM/RUBIN 1983a).

$$Y_i(1), Y_i(0) \perp\!\!\!\perp T_i \tag{4.4}$$

This assumption states a unit's assignment to a treatment is independent of that unit's potential outcomes, Y_i, and that treatment assignment is, therefore, ignorable. Causal claims borne from data that satisfy this requirement are regarded as unbiased as all confounding factors that could induce a dependence between Y_i and T_i are equally represented in the treatment and comparator arms (RUBIN 1974). Consequently, this means that the distribution of features is the same in both arms and features are said to be *balanced*. Other assumptions, such as positivity and the Stable Unit Treatment Value Assumption (SUTVA), are also necessary and assumed to be true (RUBIN

1980).

Matching and weighting are popular pre-analysis manipulations to approximate the unconditional form of strong ignorability in observational populations. These methods create pseudo-populations in which the assumption is met without need for further manipulation (Rubin 1973b). This is opposed to methods of statistical adjustment, which occur peri-analysis, and approximate the conditional form of strong ignorability (Leger 1994). Arguably, the most common strategy for weighting is the *inverse probability of treatment weighting* (IPW) (Thoemmes/Ong 2016), though other methods include the direct minimization of imbalance (Gretton et al. 2009; Kallus 2016, 2017) or weighting by the odds of treatment, kernel weighting, and overlap weighting (Rosenblatt 1956; Hellerstein/Imbens 1999; Hazlett 2016; Li et al. 2018; Kallus 2018b).

A commonality among these methods is that they implicitly or explicitly all specify a distribution function, P, that the expectation in Eq. 4.2 is taken with respect to. This distribution is often the distribution associated with the treated ($p_1(x)$), the controls ($p_2(x)$), or a combination thereof (e.g. $\frac{1}{2}p_1(x) + \frac{1}{2}p_2(x)$). This choice of distribution can lead to high variance effect estimates in circumstances where there are regions of poor overlapping support between the treated and untreated populations. An effect of this is often observed in the context of IPW analyses with instability due to propensity scores near zero or one (Kang/Schafer 2007).

In this work, we instead construct an implicit distribution, P, that focuses on the regions of the sample space with significant overlap between the treated and untreated populations. Such a construction involves an inherent trade-off between coverage and variance. For example, mixture distributions that will be valid for a larger region of the sample space will also produce high variance estimates in the context of a fixed sample budget. In the context of infinite sample sizes and positivity, one could specify any distribution P without concern for effect estimate variance. The mixture distribution of the treated and untreated populations would be a reasonable choice given a goal of maximizing coverage. However, in real-world settings with limited data, positivity may not be present and ATE estimates over such a distribution may be high variance in practice and theoretically invalid. In such a setting, valid estimates can only be made for subpopulations with significant distributional overlap. We formulate an approach that constructs a distribution P for estimating an ATE that both maximizes coverage and minimizes variance. Informally, P can be considered the distribution of a *natural experiment* where the choice of treatment, T, is independent of potential confounders, X.

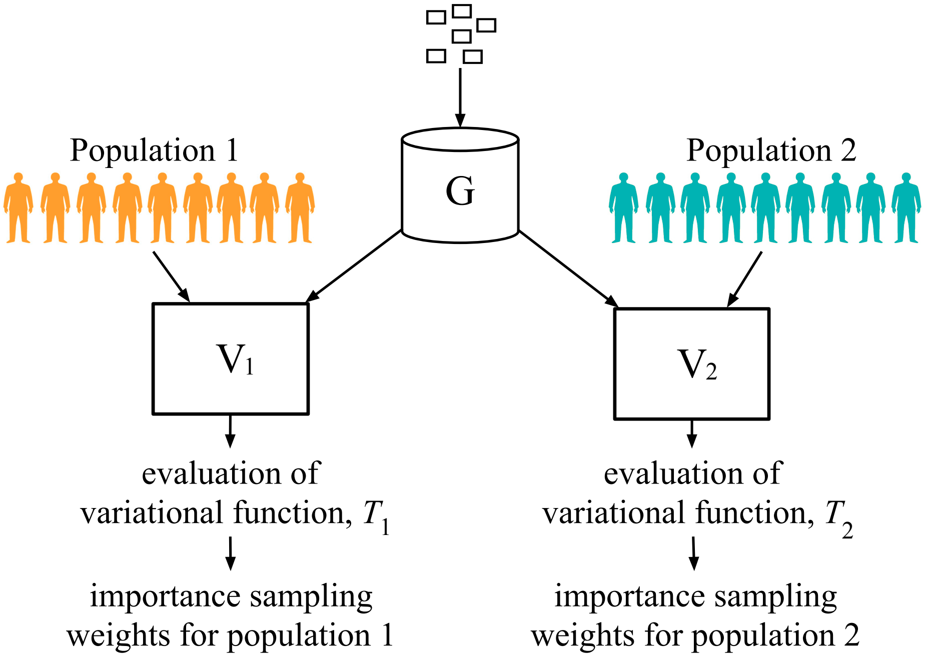
We propose the Counterfactual χ-GAN (cGAN) that uses an adversarial approach to learn a distribution that trades off coverage and effect estimate variance for two or more observational study arms. This approach learns stable, feature balancing weights without reliance on the propensity score. The target distribution, P, is identified by minimizing the Pearson χ^2-divergence between P and the sampling distributions Q_a for each study arm. To our knowledge, this is the first such application of the

Pearson χ^2-divergence. Because P is being compared to all study arms, this encourages coverage, while, as we will show, the χ-divergence inherently minimizes the variance of importance sampling estimates of the ATE.

Research Questions. *Can a generative adversarial network (GAN) be leveraged to learn feature-balancing weights?*

The Model We introduce the *Counterfactual* χ*-GAN* (cGAN), an adversarial approach to feature balance in causal inference that is based on importance sampling theory. Using an adversarial approach based on variational minimization based on the f-GAN, we minimize the sum of the Pearson χ^2-divergences between a deep generative model and the sampling distributions from each arm of a study. We show that minimizing the χ^2-divergence is equivalent, up to a constant factor, to minimizing the variance of importance sampling estimates to be made in approximating quantities such as ATEs. Similar to other weighting approaches, this approach assumes SUTVA, positivity, and no unmeasured confounders. In the following, P is the constructed target distribution and Q_a is the sampling distribution for each study arm.

Importance Sampling and the χ^2-divergence Importance sampling is a strategy for estimating expectations under an unknown target distribution given a known proposal distribution (Muller 1966). Though the importance sampling has broader usage than our application, we focused on the use of importance sampling for estimation of the average treatment effect (ATE) because of its close relationship with the χ^2 divergence. The importance sampling weight is defined as a likelihood ratio: the



Population 1
G
Population 2
V1
V2
evaluation of variational function, T_1
evaluation of variational function, T_2
importance sampling weights for population 1
importance sampling weights for population 2

Figure 4.1: Architecture of Counterfactual χ-GAN

likelihood of an observation under the target distribution, $p(x)$ divided by the likelihood under the proposal distribution, $q(x)$. Weighted expectations based on the proposal distribution approximate unweighted expectations from the target distribution at shown in Eq. 4.5.

$$\mathbb{E}_q \left[\frac{p(x)}{q(x)} \phi(x) \right] = \mathbb{E}_p \left[\phi(x) \right] \tag{4.5}$$

Consider the units in an arm of an observational study as being samples from such a proposal distribution. One strategy for obtaining unbiased expectations of treatment effects is to identify importance sampling weights for each arm that approximate expectations from a shared target distribution. However, this problem is underspecified given that we could choose any target distribution with the correct support. In this work, we choose the target distribution that yields importance sampling approximations with smallest variance. Eq. 4.6 shows the form for the variance of importance sampling estimates where $\phi(x)$ is the constant function. This choice is to make the formulation of the cGAN as outcome agnostic as possible. This form highlights its connection with the χ^2-divergence, which has a function form as shown in Eq. 4.7. This connection was previously noted in (Dieng et al. 2016). Therefore, the solution which minimizes the χ^2-divergence would also minimize the variance expectations for unknown outcomes. Of note, importance sampling is known to be a method that can produce high variance estimates, but since we will be minimizing the variance directly, this is less of a concern here.

$$\sigma_q^2 = \frac{\mu^2}{n} \left(\int q(x) \left[\frac{p(x)^2}{q(x)^2} - 1 \right] \right) dx \tag{4.6}$$

$$\chi^2(p \parallel q) = \int q(x) \left[\frac{p(x)^2}{q(x)^2} - 1 \right] dx \tag{4.7}$$

Likelihood Ratios, Overlap, & the ATE Importance sampling weights can be leveraged to estimate an ATE in that region of $q(x)$ where there is significant overlap of probability mass/density between treatment arms. This is the region that satisfies the idea of a natural experiment and in which ATE estimations are reliable. Informally, we seek to get the most coverage of the overlapping region of $q(x)$, as it results in importance sampling estimates with low variance.

Typically, the expectation in the ATE is taken with respect to the original feature distribution, $q(x)$. Under cGAN-weighted data, expectations are taken with respect to the target distribution $p(x)$. As such, calculations of the ATE from the cGAN are not equivalent to what many would classically consider the ATE, but rather, is an ATE with respect to the new, learned feature distribution. We call this new estimate the ATE_p. This inequality is demonstrated in Equation 4.8. This set of equations shows that the typical ATE, ATE_q, is not equivalent to the expectation that we estimate, the ATE_p.

$$\begin{aligned}
ATE_q &= \mathbb{E}_{q(y_1)}[y_1] - \mathbb{E}_{q(y_0)}[y_0] \\
&= \mathbb{E}_{q(x)}\mathbb{E}_{q(y_1|x)}[y_1|x] - \mathbb{E}_{q(x)}\mathbb{E}_{q(y_0|x)}[y_0|x] \\
&= \mathbb{E}_{q(x)}\mathbb{E}_{q(y|x,t=1)}[y|x,t=1] - \mathbb{E}_{q(x)}\mathbb{E}_{q(y|x,t=0)}[y|x,t=0] \\
&= \mathbb{E}_{q(x|t=1)}\frac{q(x)}{q(x|t=1)}\mathbb{E}_{q(y|x,t=1)}[y|x,t=1] - \mathbb{E}_{q(x|t=0)}\frac{q(x)}{q(x|t=0)}\mathbb{E}_{q(y|x,t=0)}[y|x,t=0] \\
&\neq \mathbb{E}_{q(x|t=1)}\frac{\mathbf{p}(x)}{q(x|t=1)}\mathbb{E}_{q(y|x,t=1)}[y|x,t=1] - \mathbb{E}_{q(x|t=0)}\frac{\mathbf{p}(x)}{q(x|t=0)}\mathbb{E}_{q(y|x,t=0)}[y|x,t=0]
\end{aligned} \tag{4.8}$$

Consider two distributions Q_1 and Q_2 that represent two arms of a study. It is possible to make unbiased ATE_p estimates based on a single distribution, P, leveraging likelihood ratios/importance sampling weights as shown in Eq. 4.9.

$$ATE_p = \mathbb{E}_p[Y_1] - \mathbb{E}_p[Y_0] = \mathbb{E}_{q_1}\left[\frac{p(x)}{q_1(x)}Y_1\right] - \mathbb{E}_{q_2}\left[\frac{p(x)}{q_2(x)}Y_0\right] \tag{4.9}$$

We will leverage an approach based on adversarial learning to simultaneously maximizes coverage, minimizes the variance defined in Eq. 4.6, and directly estimates likelihood ratios, $\frac{p(x)}{q_1(x)}$ and $\frac{p(x)}{q_2(x)}$.

f-GAN The f-GAN framework provides a strategy for estimation and minimization of arbitrary f-divergences based on a variational divergence minimization approach (Nowozin et al. 2016).

$$D_f(P \parallel Q) = \int_{\mathcal{X}} q(x) \sup_{t \in dom_{f^*}} \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\} dx \tag{4.10}$$

$$\geq \sup_{T \in \mathcal{T}} \left(\int_{\mathcal{X}} p(x)T(x)dx - \int_{\mathcal{X}} q(x)f^*(T(x))dx \right) \tag{4.11}$$

$$= \sup_{T \in \mathcal{T}} \left(\mathbb{E}_{x \sim P}[T(x)] - \mathbb{E}_{x \sim Q}[f^*(T(x))] \right) \tag{4.12}$$

where T is a class of function such that $T : \mathcal{X} \to \mathbb{R}$, f is the function that characterizes the χ^2-divergence, $f(u) = (u-1)^2$, f^* is the Fenchel conjugate of f, $f^*(t) = \frac{1}{4}t^2 + t$, and P and Q are probability distributions with continuous densities, $p(x)$ and $q(x)$. T is typically a multi-layer neural network. This formulation lower bounds the χ^2-divergence based on functions T, P, and Q in such a way that unbiased noisy gradients of the lower bound can be easily obtained based on samples from P and Q. In addition, the variational function, T, has a tight bound for $T^* = f'\left(\frac{p(x)}{q(x)}\right)$ which is equivalent to $2\left(\frac{p(x)}{q(x)} - 1\right)$ in the case of the χ^2-divergence. To respect the bounds of T that result in valid likelihood ratios, we represent T as a nonlinear transformation of an unbounded function V: $T(x) = g_f(V(x)) = -2 + log(1 + e^{V(x)})$. The likelihood ratio, $\frac{p}{q}$, is easily derived from here and provides the importance sampling weights necessary for approximating expectations under $p(x)$ as shown in Eq. 4.5.

The Counterfactual χ-GAN The cGAN builds on importance sampling theory and extends the f-GAN framework to learn feature balancing weights through an adversarial training process. Previously, (Tao et al. 2018) have explored importance weights from critics of divergence-based GAN models. However, unlike this method and other f-GANs where there is a generator, G and a single variational function, the cGAN employs dual training from at least two variational functions (Figure 4.1).

Consider a set of A treatments, each associated with one of A populations, or arms of a study. Each population contains N_a units and are drawn from an unknown and population-specific distribution Q_a. Based on the connection between the χ^2-divergence and the variance of importance sampling estimates outlined above, our objective is to identify a target distribution that minimizes the χ^2-divergence to all populations being compared: $\arg\min_p \sum_{a=1}^{A} \chi^2\left(p(x) \parallel q_a(x)\right)$. This is the sum of the divergences between the generator and the unweighted treatment arms. It is minimized when $p(x)$ equals $q_a(x)$ for all a and is directly proportional to the sum of the variances of importance sampling estimates under the target distribution, P, with proposals, Q_a. Because of the constant in Eq. 4.6, minimizing the χ^2-divergence is equivalent to minimizing a normalized variance which weighs each population equally regardless of the number of units and the magnitude of the treatment effect, ϕ.

As a byproduct of minimizing this divergence, we will also identify a set of *importance weights*, $w_{a,n}$, for each unit in each population that allows estimation of expectations from the same target distribution, P, thus satisfying the unconditional form of strong ignorability. Using these importance weights, expectations can be approximated

Algorithm 1: Minibatch stochastic gradient cGAN optimization

```
Input  : $(x_{1,1},...,x_{1,N_1},...,x_{A,N_A})$
Output : $\theta$, $\omega_{1:A}$
Initialize $\theta$, $\omega_{1:A}$ and minibatch size, $M$.
while $F(\theta, \omega_{1:A})$ not converged do
    for $a \in (1, \ldots, A)$ treatment groups do
        Sample a batch of noise samples, $z_{1:M} \sim p_g$, where $p_g$ is a prior
          distribution such as an isotropic Gaussian
        Sample minibatch of data, $x_{a,1:M} \sim q_a$
        Compute gradient w.r.t. variational function parameters
        $\nabla_{\omega_a} F = \sum_{m=1}^{M} \nabla_{\omega_a}(g_f(V_{\omega_a}(G_\theta(z_m))) - \frac{1}{4} g_f(V_{\omega_a}(x_{a,m}))^2 - g_f(V_{\omega_a}(x_{a,m})))$

        Ascend the $\omega_a$ gradient according to a gradient-based optimizer
    end
    Compute gradient w.r.t. generator parameters

            $\nabla_\theta F = \sum_{m=1}^{M} \sum_{a=1}^{A} \nabla_\theta \left[ g_f(V_{\omega_a}(G_\theta(z_m))) \right]$

    Descend the $\theta$ gradient according to a gradient-based optimizer
    Update $V_{\omega_a}$ and $G_\theta$ learning rates according to schedule
end
```

as $\mathbb{E}_p[\phi] \approx \sum_{n=1}^{N_a} w_{a,n} \phi(x_{a,n})$ where $w_{a,n} = \frac{1}{c} \frac{p(x_{a,n})}{q_a(x_{a,n})}$, where $c = \sum_{n=1}^{N_a} \frac{p(x_n)}{q_a(x_n)}$ is n normalizing constant, p is the density of the shared target distribution, q_a is the density of the proposal distribution, and $x_{a,n} \sim Q_a$. Note that our strategy eliminates the need to explicitly evaluate $p\,(x_{a,n})$ and $q_a(x_{a,n})$ as the likelihood ratio is estimated directly by the f-GAN. If desired, expectations can also be approximated using the sample-importance-resampling (SIR) algorithm where samples approximately distributed according to p can be simulated by drawing samples from the weighted empirical distribution $\hat{q}_a(x) = \frac{1}{N_a} \sum_{n=1}^{N_a} w_{a,n} \delta(x - x_{a,n})$ (Doucet et al. 2001).

The objective function for the cGAN is shown in Eq. 4.13 and is closely related to the objective defined in (Nowozin et al. 2016). θ parameterizes the generative model and ω_a parameterizes the variational model for each treatment arm, a. In our experiments,

V_{ω_a} for all a are neural networks that mirror discriminators in the traditional GAN framework and P_θ is a neural networks that mirrors the generator. Note that the generator in the original f-GAN framework is usually Q_a. In our case, to achieve the desired directionality of the χ^2-divergence, the empirical distribution must be Q_a and the generator must be P.

$$F(\theta, \omega_{1:A}) = \sum_{t=1}^{A} \Bigg(\mathbb{E}_{x \sim P_\theta}\big[g_f(V_{\omega_t}(x))\big] + \mathbb{E}_{x \sim Q_a}\big[-\frac{1}{4} g_f(V_{\omega_a}(x))^2 - g_f(V_{\omega_a}(x))\big] \Bigg) \tag{4.13}$$

Importance weights can be computed based on the fact that the bound in Eq. 4.12 is tight for $T^*(x) = f'\left(\frac{p(x)}{q(x)}\right)$ where $f(u) = (u-1)^2$. We can therefore, approximate the desired importance weights as described in Eq. 4.5 as $w_{a,n} = \frac{g_f(V_{\omega_a}(x_{a,n}))}{2} + 1$ for all $a \in (1, \dots, A)$ and $n \in (1, \dots, N_a)$. Ultimately, the ATE can be estimated between any two treatment arms according to Eq. 4.9. For example, the ATE between arms 1 and 2 could be estimated as $A\hat{T}E = \sum_{n=1}^{N_1}[w_{1,n}Y_{1,n}] - \sum_{n=1}^{N_2}[w_{2,n}Y_{2,n}]$.

Practical Considerations In the original GAN and f-GAN formulations the gradients for the generator is replaced with a related gradient that significantly speeds convergence of the model. Because our objective is minimization of the true χ^2-divergence rather than perfect distributional matching, we do not employ this loss function trick but instead apply the gradient as derived from the loss function in Eq. 4.13. Although it is the case that the domain of the Fenchel conjugate for the

χ^2-divergence is $\mathbb{R}$, we constrained it to $t \geq -2$ which produces valid likelihood ratios.

Gradient descent-based optimization of GANs is a notedly difficult task. (MESCHEDER et al. 2018; ARJOVSKY/BOTTOU; GULRAJANI et al.) Though many methods are proposed to stabilize training, we have found it sufficient to employ a set of algorithmic heuristics: (i) standardization of our data by the joint mean and variance over all A populations prior to training; (ii) periodically re-centering the distribution of each discriminator to a noisy estimate of the mean of the generator distribution. This re-centering is accomplished by setting the value of a vector that is added to the input of the discriminators.

The approach for minibatch stochastic gradient descent for the cGAN is shown in Algorithm 1. The objective function F (Eq. 4.13) is optimized by minimizing with respect to the parameters θ of the generator and maximizing with respect to the parameters $\omega_{1:A}$ of the discriminators.

Related Work Causal inference with observational data has a rich literature that cuts across many disciplines (THRUSFIELD 2017; RUBIN 1973b, 1974; PEARL 2000) including machine learning (JOHANSSON et al. 2016; KALLUS 2018a; SHALIT et al. 2017; RATKOVIC 2014; SCHWAB et al. 2018). More specifically there have been several approaches to applying adversarial networks for counterfactual inference (KALLUS 2018a; YOON et al. 2018). However, most existing methods for counterfactual inference are not directly comparable to the cGAN, as we aim to identify the most appropriate counterfactual distribution given the available data and maximize feature balance whereas most methods evaluate ATE estimation or ITE estimation directly.

In contrast to representational learning approaches and some GAN approaches, our approach does not rely on a predefined outcome to identify matched cohorts. The approach outlined in (Kallus 2018a) is the most similar in spirit to our approach but differs in that our objective directly minimizes the variance of expectations that might be used in ATE estimation, whereas (Kallus 2018a) minimizes a bound on the variance of the average treatment effect on the treated. As a result, there is no need for a regularizer, to perform cross-validation to select an appropriate level of regularization, or perform a constrained optimization over weights.

Experimentation. To evaluate the cGAN when the ground truth is known, we applied the model on simulated data of two populations/treatment arms, $A = 2$. See Figure 4.2 Each population was comprised of two subpopulations. Each subpopulation contained 10 features, drawn from a randomly generated multivariate normal distribution with a normal-Wishart prior distribution. Population 1 was composed of an equal number of samples (N=1000) from subpopulation A and subpopulation B; and Population 2 was composed of an equal number of samples from subpopulation A and subpopulation C (N=2000). By construction, subpopulation A is a latent population associated with a natural experiment, since it is part of both Population 1 and 2.

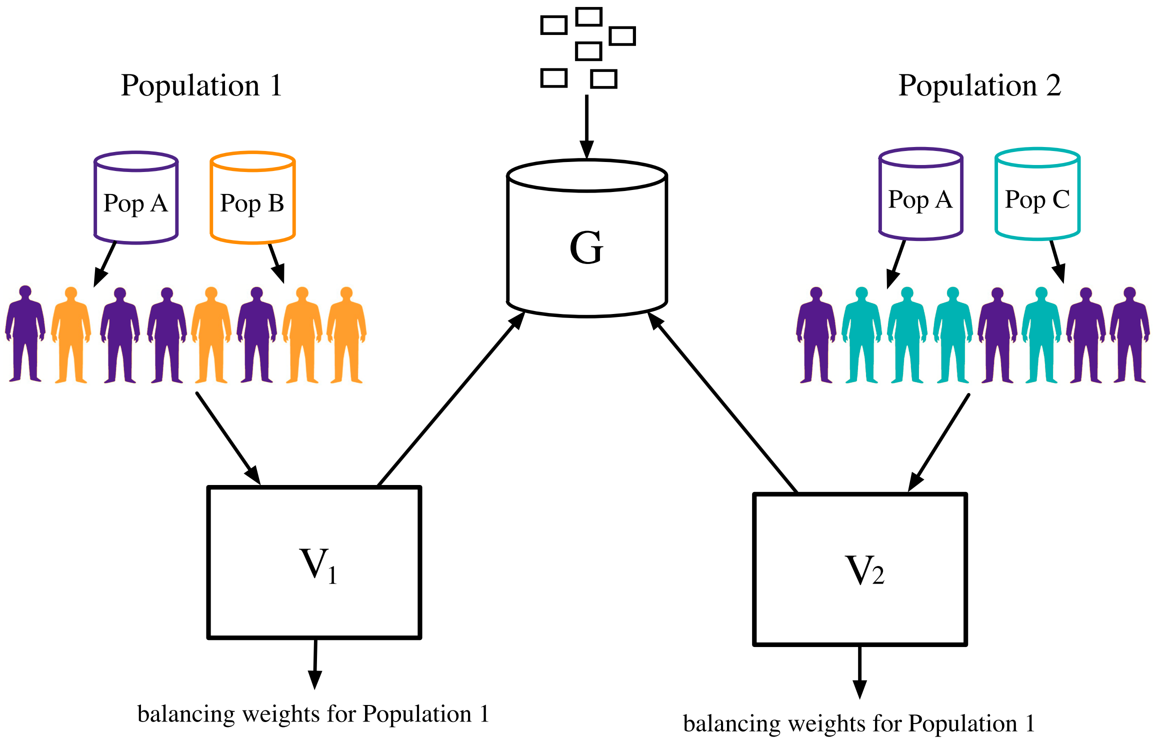

Population 1
Pop A
Pop B
G
Population 2
Pop A
Pop C
V1
V2
balancing weights for Population 1
balancing weights for Population 1

Figure 4.2: Schematic of Simulation

Because our simulation deliberately constructs populations from a shared subpopulation distribution (A), we would expect points generated from this subpopulation to have higher weights. Intuitively, the variance of importance sampling estimates should be small for both treatment groups ($a = 1$ and $a = 2$) if the learned target distribution, P_θ is one that overlaps both populations maximally while excluding density unique to one group.

To better demonstrate how the cGAN supports counterfactual reasoning, we have additionally conducted an analysis of the average treatment effect (ATE) for our experiment with simulated data. We simulated a continuous outcome according to the subpopulation of origin – Pop 1A $\sim \mathcal{N}(60, 1)$; Pop 1B $\sim \mathcal{N}(40, 1)$; Pop 2A

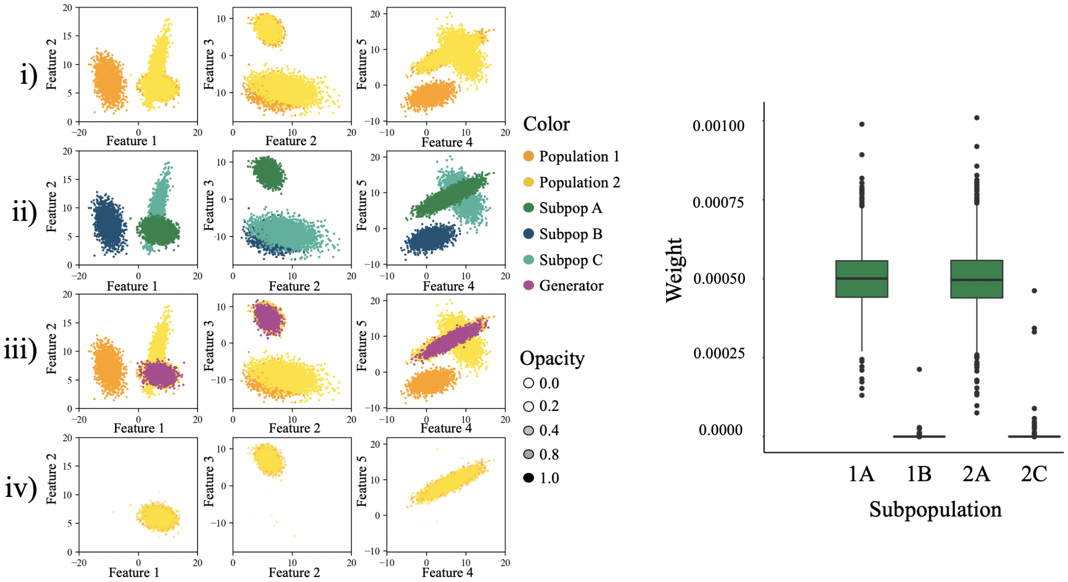


Figure 4.3: Simulation Results. *Left:* Select features (i) by population of origin; (ii) with subpopulation A highlighted; (iii) samples from the generator; (iv) opacity adjusted by weight. *Right:* Weights by subpopulation

$\sim \mathcal{N}(-10, 1)$; Pop 2C $\sim \mathcal{N}(10, 1)$. Under this outcome function, the estimate of average treatment effect (ATE) under the mixture distribution (of Pop 1 and Pop 2) is 50. When estimating the ATE under the overlapping subpopulation distribution – those from Pop 1A and Pop 2A – the ATE is 70. We applied weights from the cGAN and comparators to the simulated outcomes to assess the ability of the weighting methods to estimate one of the two ATEs. In addition, we also calculated the effective sample size (ESS), n_{eff}, using the Kish Method (Kish 1965). The ESS may be used to determine the quality of a Monte Carlo approximations of importance sampling. The calculation of n_{eff} can be found in the equation below, wherein w are the weights.

$$n_{eff} = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}$$

To investigate (i) feature-balancing weights, (ii) the biasedness of ATE, and (iii) the ESS, a variety of comparator methods were implemented in addition to the cGAN . They include binary regression propensity score; generalized boosted modeling of propensity scores (McCaffrey et al. 2004b); covariate-balancing propensity scores (Imai 2013); non-parametric covariate-balancing propensity scores (Fong et al. 2018); entropy balancing weights (Hainmueller 2011); empirical balancing calibration weights (Chan et al. 2016); optimization-based weights (Keele/Zubizarreta 2014). To better understand how simulation parameters effect cGAN and comparator performance on ATE and ESS, we have additionally implemented a sensitivity analysis. This sensitivity analysis explores how combinations of (i) the per-arm sample size (N); (ii) the unbiased average treatment effect that exists in the truly counterfactual populations ('true' ATE); and (iii) the size of the truly counterfactual populations as a proportion of the total population (overlap) effect the outcome measures. In addition to the simulation parameters outlines above – which outlines a per-arm population size of 2000 in which the size of the truly counterfactual populations is 0.5 (50%) of the population, and an unbiased, 'true' ATE of 50 – simulations were replicated for all combinations of N=[2000, 4000, 8000], overlap=[0.1, 0.5, 0.9] and a 'true' ATE = [400, 70, 0.2]. This range for the sensitivity analysis represents the breadth of values that may be present in these parameters.

Results. The results of our simulation is summarized in Figures 4.3. In the left hand-side of the Figure, the columns show the marginals of three pairs of continuous features. Row (i) shows the raw data, colored by which population units were drawn from. Row (ii) shows the same data as above, but coloring by subpopulation to highlight the overlapping distribution. Row (iii) shows a set of samples from the generator after training colored in blue. Row (iv) depicts the original data from Row (i) with the opacity of data points reflecting the importance weights. The right-hand side of the Figure shows the distribution of weights by subpopulation. Note that, in both Populations 1 and 2, the mean weights of units from subpopulation A have weights near $5x10^{-4}$, which is the uniform weight when 2000 units are in each population. Units from other subpopulations have near negligible weights, and would not meaningfully contribute to expectations in 4.9.

Subpopulation	Mean Weight
1A	$4.997x10^{-4}$
2B	$2.557x10^{-7}$
2A	$4.992x10^{-4}$
2C	$7.863x10^{-7}$

Table 4.1: Results of Application to Simulated Data. Mean cGAN-weight by subpopulation.

In the left-most figure, as you move down any column of feature pairs, it is apparent that points from the overlapping subpopulation A are both captured by the generator and assigned higher weights. This is confirmed by plotting the weights of data points by subpopulation (right-hand side of 4.3). Weights from subpopulations 1A and 2A are substantially higher than those from subpopulations 1B and 2C.

The results of this simulation further demonstrate that the ATE estimate from cGAN-weighted data is less biased than estimates from other weighting methods, given their respective targets. By construction, the causal effect of the comparable subpopulations is 70. cGAN-weighted data produced an ATE of 70.01. We see similarly good performance when inspecting the ESS. The cGAN has an ESS of 3870. Given that there are 4000 units that are comparable across the two arms (each subpopulation contains 2000 units), this is an appropriate estimate (Table 4.2).

The results of the sensitivity analysis can be found in Appendix for Aim 3.1. The results of this analysis show that superiority of cGAN performance over comparator methods persists despite the simulations parameters. Across all combinations of per-arm sample size, overlap, and 'true ATE, the cGAN consistently produced the least biased estimate of ATE and yielded the maximally appropriate ESS given the parameters.

Weighting Method	**ATE**	**ESS**
unweighted	50.03	8000
IPW	92.00	6551
clipped IPW	87.24	6997
binary regression PS	92.00	6551
generalized boosted modeling PS	84.51	7207
covariate balancing PS	91.83	6686
non-parametric covariate balancing PS	37.65	11
entropy balancing	104.13	65
empirical balancing calibration weights	52.06	65
optimization-based weights	52.07	114
cGAN	**70.01**	**3870**

Table 4.2: Results of Simulation. The average treatment effect and effective sample size (ESS) after application of weighting methods from the Counterfactual χ-GAN and comparators.

Discussion. In Aim 2.1, we introduce the Counterfactual χ-GAN. It is a deep generative model for feature balance that minimizes the variance of importance sampling estimates of treatment effects. We leverage the f-GAN framework for estimating the χ^2-divergence and likelihood ratios necessary for achieving this.

The experiments presented here suggest that cGAN is an effective method of learning feature balancing weights to support counterfactual inference. If we assume that all potentially confounding variables are observed, the superiority of cGAN in learning balancing weights, suggests that ATE borne from cGAN-weighted cohorts would be less biased than those estimates generated from traditional weighting methods.

Limitations. This method does, however, come with limitations. Training of the model is completed via backpropogation. Therefore, matching based on a combination of discrete and continuous values poses a challenge. In addition, GANs are well known for their instability and lack of objective measures for convergence. This work shares those limitations. Furthermore, assessing model convergence is a difficult task, and at present, is only evaluated uring heuristics. The simulations presented in Aim 2.1 may also present a limitation for their research. Notably, the data generation procedure outlined here may be a simplified task for the model

Aim 2.2. Apply the Counterfactual χ-GAN to observational datasets.

Background. Observational data is a vast but imperfect source of biomedical data. The calculation of unbiased causal estimates from this source are often framed as identifying a *natural experiment.* Natural experiments are a type of observational study in which researchers do not have the ability to assign the treatment, but treatments are nonetheless assigned nearly randomly. They are most valid when they closely resemble a true experimental setting, in which treatment is randomized (Meyer 1995; Shadish et al. 2002; Academy of Medical Sciences 2007; Craig et al. 2012). Popular pre-analysis methods for approximating natural experiments include matching, in which treatment units are paired with similar comparator units based on the pre-treatment features (Wilks 1932; Cochran 1953; Greenberg 1953; Billewicz 1965; Rubin 1973a); and weighting, in which units are disproportionally considered so that the weighted expectation of features are similar across arms (Czajka et al. 1992; Robins et al. 2000; Lunceford/Davidian 2004). All weighting methods generalize matching methods, and conversely, all types of matching are special cases with discrete weights (Imai 2013). Under a matching procedure, units may go unpaired, which is inefficient and may introduce new bias (Rosenbaum/Rubin 1985b; King 2011b). Weighting is a more efficient method for identifying a natural experiment. However, under many weighting techniques, downstream estimates may be unstable. A stable weighting method to support causal inference from observational data would increase the confidence of these claims and may permit the identification of effective interventions and improve outcomes. This Aim proposes the Counterfactual χ-GAN (cGAN), a variation on a generative adversarial network that learns stable feature balancing weights for two or more treatment and comparator arms.

The cGAN mechanics, intuition, and learning procedure are described in Aim 2.1.

Research Questions. *Can a generative adversarial network (GAN)-based model improve feature-balance for noisy observational cohorts?*

Methods. To determine if the cGAN can improve feature-balance in noisy observational cohorts, we additionally applied the model to observational cohorts curated according to the indication of experiments using real-world clinical data from a large, academic medical center. For this experiment, we constructed the treatment and comparator cohorts according to the protocol and indication of a published randomized clinical trial. The experiment compares sitagliptin and glimepiride in elderly patients with Type II Diabetes Mellitus (N=144 per arm) (Hartley et al. 2015). We present the 37 most frequent clinical measurements from the electronic health record.

Data. Cohorts were created from the Columbia University Medical Center's EHR according to the indication of a published RCT. The selected trial compares sitagliptin and glimepiride in elderly patients with Type II Diabetes Mellitus (T2DM). Patients were identified from the NewYork-Presbyterian Hospital EHR according to the indication and age restriction of the published RCT. Eligible patients must have at least two diagnoses of T2DM, never had an inpatient exposure to either drug, never had a previous prescription of either drug, but had at least one new prescription of either drug. Data from 144 patients taking sitagliptin and 144 patients taking glimepiride were included.

Evaluation. We evaluate the ability of the cGAN to improve feature balance by comparing the Absolute Standardized Difference of Means (ASDM) between the treatment and comparator cohorts under different weighting methods. the ASDM is a popular method of assessing cohort similarity, with a lower metric corresponding to improved feature balance. The ASDM is presented for the cGAN and the comparator weighting methods – binary regression propensity score; generalized boosted modeling of propensity scores (McCaffrey et al. 2004b); covariate-balancing propensity scores (Imai 2013); non-parametric covariate-balancing propensity scores (Fong et al. 2018); entropy balancing weights (Hainmueller 2011); empirical balancing calibration weights (Chan et al. 2016); optimization-based weights (Keele/Zubizarreta 2014). Under the clipped-IPW procedure, propensity scores greater than 90th percentile and less than 10th percentile are assigned to the values of the percentiles at 90th and 10th, respectively (Cole/Hernán 2008).

Results. The ASDM for the clinical cohorts is presented in Figure 4.4. These findings are summarized by the mean ASDM over all features, under the varying weighting methods in Table 4.3. cGAN improved mean ASDM from the unweighted cohort and improved feature balance the most among all evaluated methods. Note that this task is

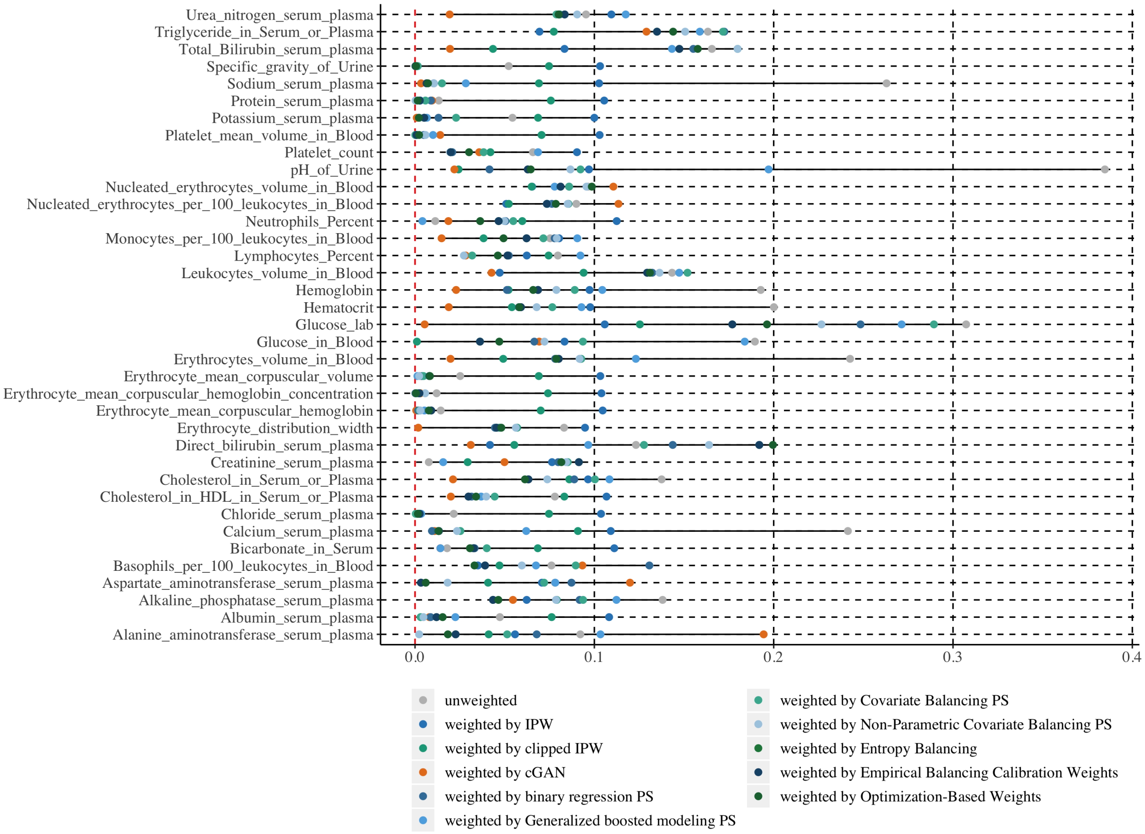


Figure 4.4: Absolute standardized difference of the means (ASDM) of real-world clinical features after application weighting methods from the Counterfactual χ-GAN and comparators.

particularly challenging due to the high dimensionality of the data and small study size.

The results of this experiment can be found in Figure 4.4 and Table 4.3. They demonstrate that cGAN-weighting achieves better feature balance than comparator methods.

Discussion. The application of the model to real-world EHR data, demonstrates that this method could provide an alternative means to causal estimation from observational data when the assumptions of no unobserved confounding, positivity,

Weighting Method	**ASDM**
unweighted	0.1103
IPW	0.0876
clipped IPW	0.0631
binary regression PS	0.0625
generalized boosted modeling PS	0.0749
covariate balancing PS	0.0681
non-parametric covariate balancing PS	0.0596
entropy balancing	0.0524
empirical balancing calibration weights	0.0524
optimization-based weights	0.0536
cGAN	**0.0364**

Table 4.3: Results of Application to Clinical Data. Absolute standardized difference of the means (ASDM) of real-world clinical features after application weighting methods from the Counterfactual χ-GAN and comparators.

and SUTVA are met. Our experiments suggest that the flexibility of our framework produces improved feature balance relevant for valid causal estimates. Furthermore, the use of cGAN-learned likelihood ratios/importance sampling weights to identify this overlapping population, permits the visualization of distributions region in which expectations – such as the ATE – are well-estimated.

Limitations. As noted in Aim 2.1, the cGAN does have limitations, including challenges in a training procedure that accommodate both continuous and discrete data types and the assessment of convergence. Though sensitivity to hyperparameters was present in the simulation of Aim 2.1, it was a greater encumbrance in the application to clinical data. The model is very response to small changes in hyperparameters, which is compounded in high-dimensional settings. The experimental set up for this Aim also suffers limitations. First and foremost, preprocessing assumptions were made to faciliate the model's application to clinical data. These include (i) which measurement to late, if multiple of the same measurements were present and (ii) imputation if a measurement was missing. Both of these techniques may affect or bias the data in some way.

In order to determine the utility of the cGAN in supporting unbiased causal claims from observational data, it would need to be compared to the gold-standard. In this case, the gold-standard would be the prospective, randomized, experiment that we based our cohort off of, HARTLEY et al. 2015. However, we found that the observational population from NYP differed dramatically from the experimental populations. As such, these populations will never be comparable, despite enforcement of exchangeability by the model. This could be addressed by multi-site collaboration, provided the patient heterogeneity matches that of the target RCT.

Chapter 5

Aim 3. Develop a method for high-throughput attributable risk estimation with observational data.

Attributable risk describes the risk for an outcome that can be allocated to a particular cause. Observational attributable risk estimation is implicitly causal inference and as with all causal inquiries with observational data, the predicating assumptions are important. In the high-dimensional setting with many potential causes, it may be difficult to construct expert-guided directed causal graphs. Instead, we develop a model that applies a simple assumption that captures an intuitive notion of causality known as causal independence. Typical methods of high-dimensional attributable risk estimation include signal detection approaches such as the Gamma Poisson Shrinker (GPS) and regression approaches such as Penalized Logistic Regression (PLR). Models such as PLR make good predictions of the outcome for individuals, but often result in global estimates that are unstable or lack interpretability as attributable risk estimates. Models such as the GPS make interpretable global estimates of risk, but are univariate, cannot account for confounding, and lack inferences for individuals. This research proposes the Noisy-Or Risk Allocation (NORA) model, a multivariate latent variable model that assumes causal independence to estimate global risks, predict outcomes, and estimate causes at the individual level. We applied NORA in simulation and to clinical data and demonstrate that it is able to predict outcomes with similar or better performance than related methods, recover known, clinically meaningful attributable risk estimates, and produce interpretable estimates of the causes for an individual's outcome.

Aim 3.1. Implement a probabilistic model, Noisy-OR Risk Allocation (NORA), and develop efficient probabilistic inference procedures.

Background. In the practice of clinical medicine, we seek to intervene on the normal course of disease to improve health. This assessment begins with the identification of the events, conditions, or characteristics – the *exposures* – that play a role in the occurrence of our outcome (effect) of interest.(Rothman 1976) The relationship between exposure and outcome is quantified by measures of association, most common of which is *risk*. In epidemiology, the risk is the is the proportion of subjects that develop the outcome of interest within a specific period, out of all subjects followed within the same period.(Cole et al. 2015) A variation of risk is the *attributable risk* (AR). AR is the proportion of an outcome in a population that could be prevented by elimination of a causal exposure from the population if there are (i) no interactions between causal exposures and (ii) all other effects of exposures are removed. (Leviton 1973; Miettinen 1974; Markush 1977) The AR may also be referred to as the population AR (Breslow/Day 1980; Boslaugh/McNutt 2008); the risk difference (Sinclair 2003), or the population etiologic fraction (Kleinbaum et al. 1982; Schlesselman/Stolley 1982).

ARs support the inference of whether a given outcome was *caused* by a particular exposure. (Rosen 1978) Such an assessment can be made at the population level (*global inference*) or at the individual level (*local inference*). At the population level, the AR can be interpreted as the proportional increase in average outcome risk over a specified time interval that would be achieved when under the exposure of interest from the population while accounting for other risk factors. (Rockhill et al. 1998; Greenland/Robins 1988) While at the individual level, the AR can be interpreted as the increase in outcome risk for a particular patient that would be achieved when under the exposure of interest, given that individuals other exposures. AR's local inferences could be used to inform treatment by highlighting the likely cause of an outcome for a single patient. While the global inferences may assist in identifying risk factors to be prioritized in public health policy, or the treatment of a single patient in which features are not known.

Ideally, estimation of AR would be based on knowledge of the relevant causal graph, in which relationships between exposures, outcomes, and confounders are made explicit. But in the setting of many potential exposures – the *high-dimensional setting* – causal graph construction may be infeasible. As an alternative, confounders to AR estimation may be controlled for through propensity-score modeling, however this may be inefficient when estimating AR for many exposures. However, this approach does not lend itself to making local inferences about a particular patient. In this work, we explore a particular model specification for estimating attributable risks in the context of unstructured binary exposures and outcomes.

In this setting, typical methods of AR estimation include the (i) calculation of excess risk by the (Levin 1953) definition; (ii) approximation using disproportionality methods for signal detection such as RRs or the Gamma Poisson Shrinker (GPS); (iii) and regression-based methods such as Penalized Logistic Regression (PLR). The Levin Definition and approximation by disproportionality methods, make interpretable global estimates of ARs, but lack inferences for individuals. Furthermore, these methods are univariate and cannot account for confounding. Regression-based methods are powerful tools for the prediction of individual-level outcomes, but may often result in global estimates that are unstable or lack interpretability as AR estimates.

This research proposes the Noisy-Or Risk Allocation (NORA) model. NORA is a multivariate latent variable model with a likelihood that captures the notion of causal independence. Unlike comparator methods, NORA is able to estimate global ARs, predict outcomes, and estimate ARs of exposures at the individual-level.

Research Questions. *Will a Bayesian model that encodes the assumption of causal independence produce attributable risk estimates that are less biased than other state-of-the-art methods?*

The Model. NORA is a Bayesian, probabilistic model that supports AR estimation of an uncertain causal system in which many potential risk factors exist. Let N be the number of patients and K be the number of unique exposures. $X_{n,k}$ is a binary indicator of exposure k for patient n; $Z_{n,k}$ is a binary indicator of activation of exposure k for patient n; R_k is the AR of exposure k; and Y_n is a binary indicator of outcome for patient n. Activation of an exposure is defined as a binary variable representing whether that exposure is a cause of the outcome for patient n. In other words, for an exposure $X_{n,k}$ to contribute to the outcome Y, it must be present ($X_{n,k} = 1$) and activated ($Z_{n,k} = 1$). The activation of the k^{th} risk for the n^{th} person is given by $Z_{n,k}$, which is dependent on both the presence ($X_{n,k}$) and the risk (R_k) of the exposure, k.

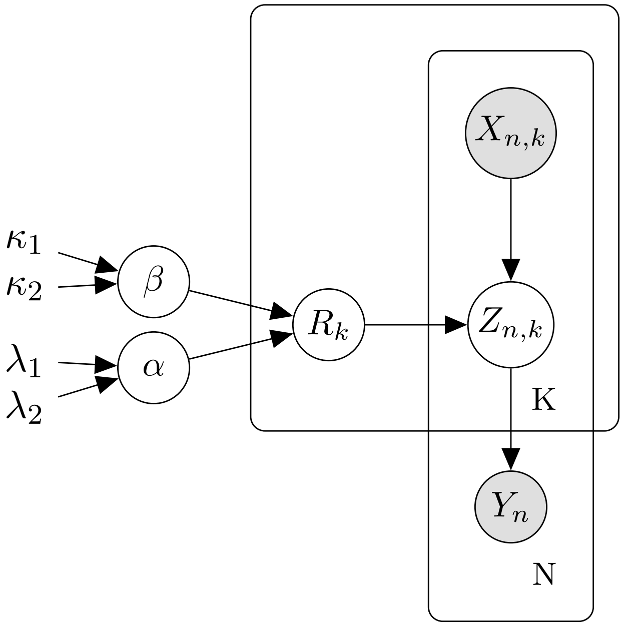


Figure 5.1: Noisy-Or Risk Allocation Model

The model is predicated upon the Noisy-Or Gate, a model which expresses the conditional probabilities of one or more binary exposures on a single binary outcome (Good 1959; Cheng 1997; Pearl 1993; Kim/Pearl). The modeling assumptions of NORA are intuitive and many causal problems of interest can be distilled into a binary representation of the data. The exposures are assumed to affect the outcome independently, a property known as *causal independence* (SRINIVAS 2013; BLUTNER). The notion of causal independence is formalized in Equation 5.1. This assumption means that the probability of surviving an outcome given an exposure, is independent of the probability of surviving, given other exposures.

$$p(Y=0|X) = \prod_{i=1}^{I} p(Y=0|X_i) \tag{5.1}$$

The Noisy-Or Gate satisfies this assumption and its functional form is shown in Equation 5.2.

$$p(Y|X) = 1 - \prod_{i=1}^{I} (1 - p(Y|X_i)) \tag{5.2}$$

The generative process for NORA is based on this and given by Algorithm 2. The full joint distribution is shown in Equation 8.1 and the associated graphical model is shown in Figure 5.1.

$$p(Z, Y, R, \alpha, \beta) = p(\alpha; \lambda) p(\beta; \kappa) \prod^{N} p(Y_n|Z_{n,1:K}) \prod^{K} p(R_k|\alpha, \beta) \prod^{K} \prod^{N} p(Z_{n,k}|X_{n,k}, R_k) \tag{5.3}$$

The likelihood of Y embodies the assumption that causal influences are independent in determining the effect. Given X_n, the posterior distribution of Z_n is deterministically 0 when $Y = 0$, and when $Y = 1$ the posterior of $Z_{n,j}$ is independent of the posterior for $Z_{n,k}$ for all j and k, given that at least one cause is activated (i.e., $Z_{n,l} = 1$).

Algorithm 2: Generative Process for the NORA Model

```
Choose $\alpha \sim Gamma(\lambda_1, \lambda_2)$
Choose $\beta \sim Gamma(\kappa_1, \kappa_2)$
for $k \in \{1, \ldots, K\}$ exposures do
    Choose an attributable risk $R_k \sim Beta(\alpha, \beta)$
end
for $n \in (1, \ldots, N)$ patients do
    for $k \in \{1, \ldots, K\}$ exposures do
        Choose an activation $Z_{n,k} \sim Bernoulli(X_{n,k} R_k)$
    end
    Choose an outcome $Y_n \sim Bernoulli\left(\left[1 - \prod_{k=1}^{K}(1 - Z_{n,k})\right]\right)$ based on the
      activations
end
```

Simply, when outcome is absent ($Y = 0$), there is an absence of explaining-away and all activations (Z) must be zero; when the outcome is present ($Y = 1$), the activations (Zs) are coupled by explaining-away, only in the absence of an activated cause. This corresponds to the intuitive assumption of causal independence. For example, taking drug B does not influence whether drug A causes the outcome.

NORA has a Bayesian formulation that confers many benefits for causal AR modeling. It can accommodate AR estimation for which there is varying amount of evidence. The model can encode our assumptions about what the data look like using informative priors. Our prior on the risks (R) bounds them between zero and one, constrains the risks in a natural way. These, in combination with the non-negativity of the risks allows for inference in high dimensional datasets with limited observations.

As is the case with other adjustment methods, like logistic regression, NORA assumes strong ignorability and adjusts for confounding through conditioning on observe confounders. However, the success of such methods to support causal estimation depends on the accuracy of the likelihood assumptions. ARs from NORA may be interpreted as *causal* if the causal independence assumption holds and there is no unobserved confounding.

Inference The risks (R) and the activations (Z) were inferred using an optimized Gibbs sampler. A Metropolis algorithm was used to perform inference on the risk priors (α and β). The posterior of Z is given by Equation 5.4, which can be normalized and sampled via enumeration.

$$p(Z_{n,k}|\ldots) \propto \Big[1-(1-Z_{n,k})\prod_{j\neq k}(1-Z_{n,j})\Big]^{Y_n}\Big[(1-Z_{n,k})\prod_{j\neq k}(1-Z_{n,j})\Big]^{(1-Y_n)}(X_{n,k}R_k)^{Z_{n,k}}(1-X_{n,k}R_k)^{1-Z_{n,k}} \tag{5.4}$$

Given that R is a conjugate prior for Z, the posterior of R is Beta distributed, $Beta(\alpha^*, \beta^*)$, where α^* is given by Equation 5.6 and β^* is given by Equation 5.7. The posteriors for α and β were sampled with a Metropolis algorithm based on the likelihood of their respective Markov blankets.

$$p(R_k|\ldots) = \frac{1}{B(\alpha^*, \beta^*)}\left[R_k^{\alpha^*-1}(1-R_k)^{\beta^*-1}\right] \tag{5.5}$$

$$\alpha^* = \alpha + \sum_{n=1}^{N} Z_{n,k} \tag{5.6}$$

$$\beta^* = \beta + \sum_{n=1}^{N} X_{n,k} - \sum_{n=1}^{N} Z_{n,k} \tag{5.7}$$

Data. Data for Aim 3A will be simulated according to Figure 5.2

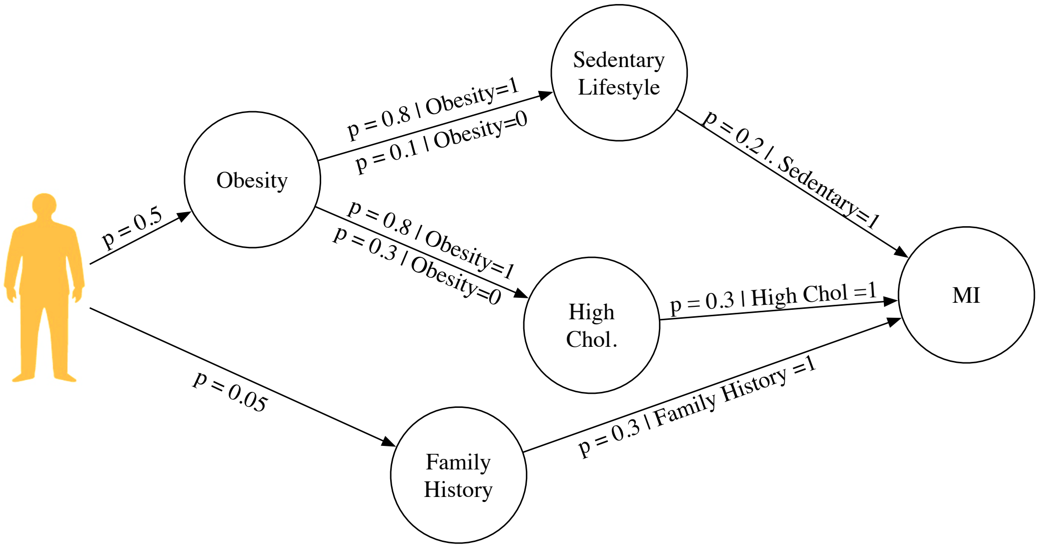


Figure 5.2: NORA Simulation Schema

Evaluation. To understand NORA's properties, we simulated data according to the toy causal system represented in Table 1, wherein High Cholesterol is the exposure with the greatest risk for the outcome, myocardial infarction (MI). We hypothesized that we would recover the risk of MI associated with High Cholesterol within a small margin of error when all variables are observed despite disregarding knowledge of the causal graph. We also hypothesize that the degree of bias due to confounding via the backdoor path would be less under the NORA model than under logistic regression (LR). The data in the simulation is generated such that Sedentary Lifestyle could serve as a confounder of the relationship between High Cholesterol and MI in the setting where Obesity and Sedentary Lifestyle are unobserved. In all simulations, we will compare the average risk estimates over 100 trials for all variables on the outcome, MI, as determined by both the NORA model and LR.

Both NORA and LR will be applied to the simulated data twice; (i) the Main Effect

Model, in which the confounding variables (Obesity and Sedentary Lifestyle) are not observed; and (ii) the Adjusted Model, in which the confounding variables are observed. The true causal attributable risk of myocardial infarction is known to be 0.3 for High Cholesterol. We evaluated the empirical bias of both methods in the simulation.

Results. NORA, modeling main-effects only, resulted in 1000-trial average for the risk of MI from High Cholesterol of 0.177; with full adjustment for confounding, NORA found the 100-trial average for attributable risk of MI from High Cholesterol to be 0.298 (Truth = 0.30); LR, when modeling main effect only found the attributable risk of MI from High Cholesterol to be 0.459; and when modeled with all confounders, resulted in a probability of 0.821. When the backdoor-path/confounders were unobserved, the risk estimate of High Cholesterol from the NORA model changed 37.1%, versus 78.9% when estimated from LR. Given the structure of our simulation, it is not surprising NORA was able to recover the true risk of High Cholesterol in our simulation, with greater accuracy than the LR. However, the misspecification of using LR when causal independence exists also led to a greater sensitivity to unobserved confounders. We also note that the lack of knowledge of the causal graph did not adversely affect the estimate despite the collinearity of High Cholesterol, Obseity, and Sedentary Lifestyle.

Discussion. NORA a new method of attributable risk estimation that supports both local and global inferences. The likelihood of the model encodes an intuitive and simplifying assumption of causality. In the absence of causal graph construction, such simplifying assumptions can be powerful tools for estimation even in the context where the assumptions are only partially met as indicated by our simulations.Our simulation demonstrates that NORA is able to recover risks with high accuracy, and indicate that it may be less prone to confounding via the backdoor path than other models when misspecified. Unlike other methods of causal inference with observational data, our simulations have shown that if there is significant collinearity in a causal graph, NORA is able to accurately estimate attributable risks nonetheless in absence of knowledge of the graph structure.

Limitations. As with all models, NORA does have limitations. The Model is Bayesian, which confers many benefits, however we make certain distributional choices and have not assessed model sensitivity to those choices. With the simulation presented in Aim 3.1, the data was generated based on an assumption of causal independence, which may be a simplified learning task for our model. However, the simulation was primarily to evaluate the impact of lack of knowledge of the causal graph and model misspecification when logistic regression is used to control for the effect of other exposures in a causally independent system.

Aim 3.2. Apply the NORA model to observational datasets.

Background. To determine NORA's utility in practice, we applied the model to ten different observational cohorts with the hypothesis that the model would be able to recover known, clinically meaningful causal relationships previously acknowledged in published literature.

We conducted several experiments, each associated with a specific outcome of interest and a set of exposures from a specific data type. Outcomes included disseminated intravascular coagulation (DIC), glaucoma, hearing loss, heart failure, Kaposi sarcoma, mucositis, renal impairment, disorder of the spleen, and hypothyroidism. More information regarding the domains of exposure can be found in Table 5.1.

NORA's mechanics, intuition, and learning procedure are described in Aim 3.1.

Outcome	Exposure domain
disseminated intravascular coagulation (DIC)	procedures
glaucoma	conditions
hearing Loss	ingredient-level drugs
heart Failure	conditions
Kaposi Sarcoma	conditions
mucositis	ingredient-level drugs
renal impairment	conditions
Disorder of the Spleen	conditions
hypothyroidism	procedures
mucositis	procedures

Table 5.1: Application to clinical data. Outcomes of interest and domains of exposures.

ARs were estimated for the causal systems within the EHR using (i) NORA, (ii) L1-regularized Logistic Regression (L1), (iii) the Levin-AR calculation (Equation 2.30), and (iv) AR estimation using DPAs, including the RR and GPS-EBGM (Equation 2.31).

Research Questions. *Can a Bayesian model that encodes the assumption of causal independence be used to support high-throughput attributable risk estimation with observational data?*

Data. Observational cohorts for NORA's application to clinical data were procured from the a large academic medical center's electronic health record (EHR) system. The EHR contains 5.4 million clinical observations from 1986 to 2017. Patients encounters are documented in the EHR at each outpatient, inpatient, and emergency department visit. Data modalities generally include, but are not limited to, diagnoses, clinical measurements, medications, and procedures.

Construction of observational cohorts was the same workflow for all outcomes. (1) Eligible patients were required to have at least 365 days of clinical observation. Because this requirement is the same for all outcomes of interest, each cohort contained the same number of patients (N=105,377). (2) From this population, patients were identified as having the outcome of interest or not. (3) The set of exposures (either procedure, condition, or ingredient-level drug) was determined as the unique set of exposures from the outcome-positive and outcome-negative patients. (i) For outcome-negative patients, this includes all exposures from the beginning to the end of the clinical record. (ii) For outcome-positive patients,this includes all exposures that occurred from the beginning of the patient record to the day before their outcome diagnosis. (4) Lastly, exposures were binarized for all patients in the cohort – for any exposure, k, $X_k = 1$ if patient n had the exposure, and $X_k = 0$ if they did not have the exposure. Details on how outcomes were defined and counts of the number of positive patients and number of unique exposures can be found in the Appendix for Aim 3.2.

To accommodate the spontaneous occurrence of the outcome in the absence of any true risk-factors, we have included an exposure that is present for all individuals (the intercept). (Henrion 1987) As the EHR may be noisy or incomplete, the intercept will permit AR estimation by NORA in the absence of perfect data from this source.

Evaluation. To contextualize the results of our experiments with clinical data, we propose a three-part evaluation that addresses (1) *the local inference of the outcome*, (2)*the global inference of the exposures*, and (3) *the local inference of the exposures.*

The *local inference of the outcome* is evaluated through the predictive performance of a held-out dataset. Cohorts were split into train and test sets (80 and 20, respectively). Such an assessment is only feasible for methods in which prediction is possible, so only NORA and L1 may be evaluated in this respect. ARs for NORA will be determined by taking the median value of model-learned risks over the final 250 iterations of inference.

The *global inference of the exposures* is evaluated by comparing the gold-standard, real-world AR estimates with the model-based AR estimates from NORA, L1, the Levin-AR calculation, RR, and MGPS. A model-based AR estimate that coincides with the real-world AR estimate is indicative of a stable, unbiased method. To evaluate this completely, thousands of exposure-outcome relationships would require a real-world AR estimates. Such a literature review is infeasible. To demonstrate how much an assessment would take place, we present an analysis of one exposure for the outcome of Kaposi sarcoma.

The *local inference of the exposures* is evaluated through an inspection of high-AR exposures for an individual with an outcome of interest. Such an assessment employs both the risk-factors that a single patient is exposed to, and the AR associated with those exposures. We evaluated the posterior distribution over the set of activations (Z_n) to evaluate the probability that a given exposure is a cause for a patient's outcome given a point estimate of the remaining latent variables. Given the large number of patients, this evaluation would be infeasible to carry out or summarize for all patients. To demonstrate how such an assessment would take place, we present an analysis of one individual for the outcome of heart failure.

Results. The exposures and ARs for the five highest risk exposures, as estimated by NORA, L1, the Levin-AR calculation, RR, and GPS-EBGM, can be found in the Appendix for Aim 3.2.

When evaluating the *local inference of the outcome*, NORA had better predictive ability than L1 for certain outcomes and L1 had better predictive performance for others. Overall, NORA has a higher average AUC of 0.6817 as compared to L1 with an AUC of 0.6669. The results of this evaluation are summarized by the area under the receiver operating curve (AUROC) reported in Table 5.2.

	NORA	L1
disseminated intravascular coagulation (DIC)	**0.8878**	0.7773
glaucoma	**0.7017**	0.6999
hearing loss	0.5056	**0.6329**
heart failure	**0.8030**	0.7953
Kaposi sarcoma	**0.8011**	0.5624
mucositis (v drugs)	0.5291	**0.6560**
renal impairment	**0.8170**	0.7965
disorder of the spleen	**0.6248**	0.5000
hypothyroidism	0.5643	**0.6349**
mucositis (v procedures)	0.5829	**0.6134**

Table 5.2: Local inference of the outcome – Area Under the Receiver Operating Curve (AUROC) for NORA (median over last 250) and L1

For each method, the three exposures with the highest ARs for Kaposi sarcoma are presented in Table 5.4. Kaposi sarcoma is a cancer of endothelial origin. Prior to the 1980s, the disease was considered rare, only occurring in a small population of men elderly men from an isolated geographic region. The disease may also arise in immuno-compromised or immuno-suppressed patients. After the 1980s, the disease increased in prevalence, with many cases due to the spread of human immunodeficiency virus (HIV) (Gupta/Kumar). A 2018 article by Liu, et al reported the AR of HIV for Kaposi sarcoma to be 0.0048. (Liu et al. 2018) NORA, L1, the Levin-AR calculation, and GPS-EBGM all identify HIV as one of the highest AR exposures. However, the estimates from L1, the Levin-AR calculation, and GPS-EBGM are extremely high, only the estimate from NORA is the correct order of magnitude. This suggests that NORA may produce accurate population-level estimates of attributable risks. The top-1 attributable risks for all evaluated outcomes are shown in Table 5 and top-5 results for each outcome can be found in Appendix for Aim 3.2.

	Gold-Standard	NORA	L1	Levin AR	RR	GPS-EBGM
Human immunodeficiency virus infection	0.0048	0.0070	0.7872	0.2022	0.9566	0.9603

Table 5.4: Global inference of the exposures – attributable risk estimates of HIV for Kaposi sarcoma vs gold-standard estimate

The results of our simulation to evaluate the *local inference of the exposures* are shown in Figure 5.3. This heart failure patient had twenty unique exposures. Of these, *cardiomyopathy* (0.367), *preinfarction syndrome* (0.084), *coronary arteriosclerosis in native artery* (0.045), and *essential hypertension* (0.023) were the highest AR exposures for this patient. These exposures are known risk-factors of heart failure and are biologically sensible. Though this patient had other exposures, such as *abdominal pain* and *sprains and strains of joints and adjacent muscles*, these do not have a causal relationship with the outcome and have an estimated AR of near zero.

Discussion. NORA a new method of attributable risk estimation that supports both local and global inferences. When applied to real-world clinical data, NORA identifies a cohesive set of high-AR risk factors that have reasonable estimates of risk. Our model is able to identify global risks for high-burden and rare clinical outcomes. To our knowledge, NORA is the only model of AR that is able to support these three types of inferences related to attributable risk estimation. By design, this model may be simultaneously applicable at the patient-level with outcome predictions and causal

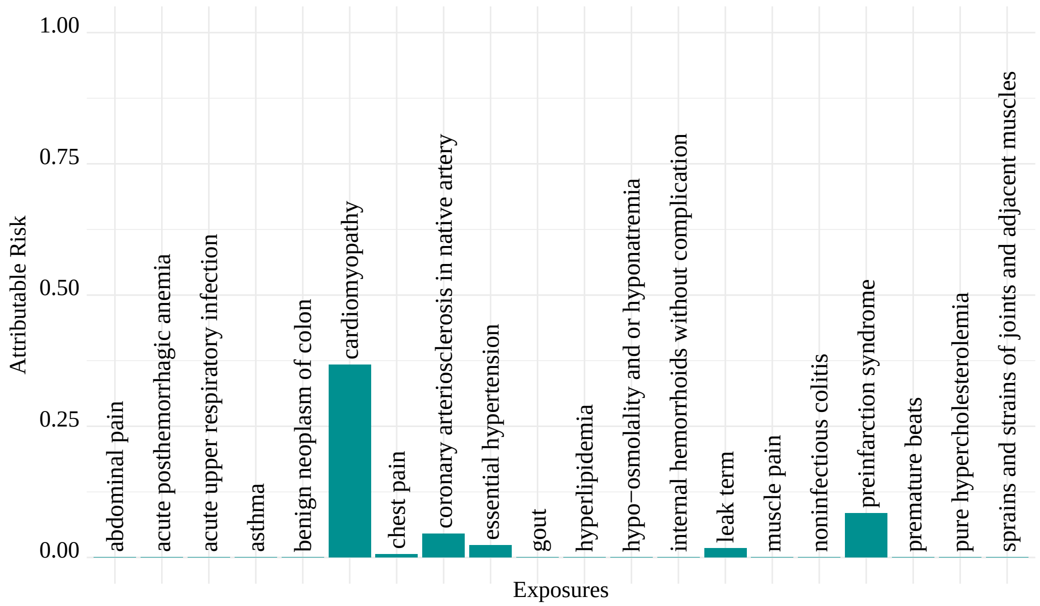


Figure 5.3: Local inference of the exposures – the average attributable risk of each exposure for a single heart failure patient.

estimation and at the population-level informing public-health with estimates of risks across the entire population.

Limitations. In addition to those noted in Aim 3.1, the application of NORA to EHR data may confer additions limitations. The model may learn artifacts of the record itself, rather than true documentation patterns. Additionally, the model depends on the accuracy of EHR timestamps. If timestamps are off then we may (i) miss certain exposures timestamp is erroneously in the future and beyond the outcome) or (ii) reverse the temporality, in which the outcome may appear to be an exposure because the timestamp is erroneously late. Lastly, the assumption of causal independence may not be true. In that context, it is unclear how the inferred parameters should be interpreted if this model is applied.

Chapter 6

Conclusions

Understanding the relationship between cause and effect answers *causal questions*, which are absolutely fundamental in many disciplines, but none more so than healthcare. In healthcare, we are interested in intervening in what would be the normal course of disease in order to improve the lives of individuals. If we understand and communicate causes causes, then we are more likely to take better action or optimize an intervention. Knowledge of what causes serious, high-burden outcomes supports the explanation and understanding of the phenomenon and informs our treatment and prevention new cases. And this is principally what happens in the practice of medicine everyday. We study things to understand what to do to live healthier lives, and we make this assessment with causal inference.

The methods that I presented in this dissertation address two complex tasks in causal inference from observational data – the replication or evaluation of *existing* causal knowledge, and the generation of *new* causal knowledge.

In Aim 1 of this dissertation, I address the ability to replicate and evaluate causal claims from observational data sources. A number of techniques exist to improve and support causal estimates from observational data, but at present, there is no widely-used framework to evaluate modeling assumptions relative to experimental data. RCTs, which we accept to be the least biased source of causal knowledge, can be compared to estimates generated from observational data and, thus, provide a methodology to assess the validity of causal claims and a platform with which to evaluate inference methods. This can serve as framework for evaluating methods for causal inference. The research in Aim 1 empirically demonstrates that eligibility criteria are not sufficient for identifying the applicable real-world population in which experimental treatment effects will replicate. For perfect replication with observational data, the distribution of treatment effect must be the same as that in the trial. Therefore, the inability of perfect replication may due attributable to the presence of HTE, which is not accounted for with the eligibility criteria. This has important consequences on the practice of EBM, as our research indicates that evidence cannot reasonably be transferred to a patient given the current data reporting standards.

In Aims 2 and 3 of this dissertation, I present new methods to generate causal knowledge from observational data. In Aim 2, I address causal knowledge that arises from the comparison of two cohorts. And in Aim 3, I address causal knowledge in the form of attributable risk estimation.

Aim 2 presents a novel, deep-learning based method that uses adversarial training to learn feature-balancing weights, called the Counterfactual χ-GAN (cGAN).

The experiments presented in Aims 2.1 and 2.2 suggest that cGAN is an effective method to learn balancing weights, that may support counterfactual inference. The application of the model to real-world EHR data, demonstrates that this method could provide an alternative means to causal estimation from observational data when other assumptions of counterfactual inference are met. Our extended simulations suggest that our framework is flexible to produce valid causal estimates from a variety of settings. Furthermore, if all confounding variables are assumed to be observed, the superiority of cGAN suggests that average treatment effects (ATE) borne from cGAN-weighted cohorts would be less biased than comparator methods.

Aim 3 presents a new, high-throughput method of AR estimation – Noisy-Or Risk Allocation (NORA) model. In all causal inquiries, the predicating assumptions are important. In the high-dimensional setting, the assumption of causal independence is particularly important because it supports inference in the presence of many causes – the multivariable setting. When the causal graph is infeasible to construct in this setting, but we would like to make inferences over many exposures, the assumptions of NORA offer advantages for unbiased AR estimation. The model combines the Bayesianism and the multivariable modeling with a likelihood that captures the notion of causal independence. This allows the model to effectively account for confounding and collinearity. assuming conditional ignorability and SUTVA, the experiments presented in Aims 3.1 and 3.2 demonstrate that NORA is able to recover known, clinically meaningful causal relationships with similar or better performance than the state-of-the-art. Furthermore, unlike comparator methods, which are unable to support inferences at the global and local levels. To our knowledge, NORA is the only model of attributable risk that is able to support both types of inference, which helps us rectify care of a single patient with public-health and aids in our understanding of health-promotion.

Chapter 7

References

Academy of Medical Sciences. 2007. *Identifying the enviornmental causes of disease: how should we decide what to believe and when to take action.* Tech. rep. London.

Ali, Ayad K. 2011. “Pharmacovigilance analysis of adverse event reports for aliskiren hemifumarate, a first-in-class direct renin inhibitor.” *Therapeutics and clinical risk management* 7:337–44.

Anderson-Cook, Christine M. 2005. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. arXiv: `NIHMS150003`.

Anglemyer, Andrew/Hacsi T Horvath/Lisa Bero. 2014. “Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials”. *Cochrane Database of Systematic Reviews*, no. 4: MR000034.

Angrist, Joshua D./Guido W. Imbens/Donald B. Rubin. 1996. “Identification of Causal Effects Using Instrumental Variables”. *Journal of the American Statistical Association* 91 (434): 444–455. arXiv: `arXiv:1404.1785v2`.

Angrist, Joshua D./Jörn-Steffen Pischke. 2010. “The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics”. *Ssrn* 24 (2): 3–30.

Angrist, Joshua/Donald B Rubin/Guido Imbens. 1993. “Identification of Causal Effects Using Instrumental Variables : Rejoinder”. *American Statistical Association* 91 (434): 444–455.

Arbogast, Patrick G./Wayne A. Ray. 2009. “Use of disease risk scores in pharmacoepidemiologic studies”. *Statistical Methods in Medical Research* 18 (1): 67–80.

Arjovsky, Martin/Léon Bottou. *TOWARDS PRINCIPLED METHODS FOR TRAINING GENERATIVE ADVERSARIAL NETWORKS*. Tech. rep. arXiv: `1701.04862v1`.

Austin, Peter C. 2011a. “An introduction to propensity score methods for reducing the effects of confounding in observational studies”. *Multivariate Behavioral Research* 46 (3): 399–424. arXiv: `arXiv:1011.1669v3`.

— . 2011b. “Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies”. *Pharmaceutical Statistics* 10 (2): 150–161.

— . 2010. “Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score”. *American Journal of Epidemiology* 172 (9): 1092–1097.

Baltar, Valéria Troncoso/Clóvis Arlindo de Sousa/Marcia Faria Westphal. 2014. “Mahalanobis’ distance and propensity score to construct a controlled matched group in a Brazilian study of health promotion and social determinants”. *Revista Brasileira de Epidemiologia* 17 (3): 668–679. arXiv: `1202.1088`.

Barnow, Burt S./And Others. 1980. *Issues in the Analysis of Selectivity Bias. Discussion Papers. Revised.*, vol. 5. San Francisco: SAGE Publications.

Bartlett, Victoria L/Sanket S Dhruva/Nilay D Shah/Patrick Ryan/Joseph S Ross. 2019. "Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence." *JAMA network open* 2 (10): e1912869.

Bate, Andrew/S. J.W. Evans. 2009. "Quantitative signal detection using spontaneous ADR reporting". *Pharmacoepidemiology and Drug Safety* 18 (6): 427–436. arXiv: `NIHMS150003`.

Becker, Sascha. 2016. "Using instrumental variables to establish causality". *IZA World of Labor* 250.

Belloni, Alexandre/Victor Chernozhukov/Christian Hansen. 2013. "Inference on treatment effects after selection among high-dimensional controls". *Review of Economic Studies* 81 (2): 608–650. arXiv: `1201.0224`.

Belsley, David A./Edwin. Kuh/Roy E. Welsch. 2004. *Regression diagnostics : identifying influential data and sources of collinearity.* 292. Wiley.

Billewicz, W. Z. 1965. "The efficiency of matched samples: an empirical investigation." *Biometrics* 21 (3): 623–644.

Blackwell, Matthew. 2013. "Observational Studies and Confounding Confounding Observational studies versus experiments".

Blutner, Reinhard. *Noisy OR* •.

Bollen, Kenneth/John T Cacioppo/Robert M Kaplan/Jon A Krosnick. 2015. *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. Tech. rep.

Boslaugh, Sarah/Louise-Anne McNutt, eds. 2008. *Encyclopedia of Epidemiology - Google Books*. 1st ed. Thousand Oaks, CA: SAGE Publications.

Breslow, N E/N E Day. 1980. "Statistical methods in cancer research. Volume I - The analysis of case-control studies." *IARC scientific publications*, no. 32: 5–338.

Britton, A./M. McKee/N. Black/K. McPherson/C. Sanderson/C. Bain. 1999. "Threats to applicability of randomised trials: Exclusions and selective participation". *Journal of Health Services Research and Policy* 4 (2): 112–121.

Brookhart, M. Alan/Sebastian Schneeweiss/Kenneth J. Rothman/Robert J. Glynn/Jerry Avorn/Til Stürmer. 2006. "Variable selection for propensity score models". *American Journal of Epidemiology* 163 (12): 1149–1156. arXiv: `NIHMS150003`.

Burns, Patricia B./Rod J. Rohrich/Kevin C. Chung. 2011. "The Levels of Evidence and Their Role in Evidence-Based Medicine". *Plastic and Reconstructive Surgery* 128 (1): 305–310. arXiv: `NIHMS150003`.

Cain, Lauren E./Michael S. Saag/Maya Petersen/Margaret T. May/Suzanne M. Ingle/Roger Logan/James M. Robins/Sophie Abgrall/Bryan E. Shepherd/Steven G. Deeks/M. John Gill/Giota Touloumi/Georgia Vourli/François Dabis/Marie-Anne Vandenhende/Peter Reiss/Ard van Sighem/Hasina Samji/Robert S. Hogg/Jan Rybniker/Caroline A. Sabin/Sophie Jose/Julia del Amo/Santiago Moreno/Benigno Rodríguez/Alessandro Cozzi-Lepri/Stephen L. Boswell/Christoph

Stephan/Santiago Pérez-Hoyos/Inma Jarrin/Jodie L. Guest/Antonella D'Arminio Monforte/Andrea Antinori/Richard Moore/Colin N.J. Campbell/Jordi Casabona/ Laurence Meyer/Rémonie Seng/Andrew N. Phillips/Heiner C. Bucher/Matthias Egger/Michael J. Mugavero/Richard Haubrich/Elvin H. Geng/Ashley Olson/ Joseph J. Eron/Sonia Napravnik/Mari M. Kitahata/Stephen E. Van Rompaey/ Ramón Teira/Amy C. Justice/Janet P. Tate/Dominique Costagliola/Jonathan A.C. Sterne/Miguel A. Hernán/Antiretroviral Therapy Cohort Collaboration, the Centers for AIDS Research Network of Integrated Clinical Systems, and the HIV-CAUSAL Collaboration. 2015. "Using observational data to emulate a randomized trial of dynamic treatment-switching strategies: an application to antiretroviral therapy". *International Journal of Epidemiology* 45 (6): dyv295.

Calder, Bobby J./Lynn W. Phillips/Alice M. Tybout. 1982. "The Concept of External Validity". *Journal of Consumer Research* 9 (3): 240.

Campbell, Donald T/Julian C Stanley. 1963. *Experimental and Quasi-Experimental Design for Research*, 1–84. Boston: Houghton Mifflin Company. arXiv: `arXiv:1011.1669v3`.

Canida, T. 2017. "An R Implementation of the Gamma-Poisson Shrinker Data Mining Model". *The R Journal* 9 (2): 499–519.

Cannon, Christopher P./Eugene Braunwald/Carolyn H. McCabe/Daniel J. Rader/ Jean L. Rouleau/Rene Belder/Steven V. Joyal/Karen A. Hill/Marc A. Pfeffer/ Allan M. Skene. 2004. "Intensive versus Moderate Lipid Lowering with Statins after Acute Coronary Syndromes". *New England Journal of Medicine* 350 (15): 1495–1504. arXiv: `arXiv:1011.1669v3`.

Cavuto, S./F. Bravi/M. C. Grassi/Giovanni Apolone. 2006. "Propensity score for the analysis of observational data: An introduction and an illustrative example". *Drug Development Research* 67 (3): 208–216.

Cha, Sung-hyuk. 2007. "Comprehensive survey on distance/similarity measures between probability density functions". *City* 1 (2): 1. arXiv: `0500581 [submit]`.

Chan, K C G/S C P Yam/Z Zhang. 2016. *Globally Efficient Nonparametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting.* Tech. rep. University of Washington.

Charles. 2013. "Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals : 2008-2012", no. 9: 2008–2012.

Cheng, Patricia W. 1997. *From Covariation to Causation: A Causal Power Theory.* Tech. rep. 2.

Cochran, W. G. 1953. "Matching in analytical studies." *American journal of public health* 43 (6 :1): 684–691.

Cochran, W G. 1968. "The effectiveness of adjustment by subclassification in removing bias in observational studies." *Biometrics* 24 (2): 295–313. arXiv: `9809069v1 [arXiv:gr-qc]`.

Cochran, W G/G. M. Cox. 1950. *Experimental Designs.* New York: John Wiley & Sons, Ltd.

Cole, S. R./M. G. Hudgens/M. A. Brookhart/D. Westreich. 2015. “Risk”. *American Journal of Epidemiology* 181 (4): 246–250.

Cole, Stephen R/Miguel A Hernán. 2008. “Constructing inverse probability weights for marginal structural models.” *American journal of epidemiology* 168 (6): 656–64.

Concato, John. 2004. “Observational Versus Experimental Studies: What’s the Evidence for a Hierarchy?” *NeuroRx* 1 (3): 341–347.

Contopoulos-Ioannidis, D. G./G. A. Alexiou/T. C. Gouvias/J. P. A. Ioannidis. 2008. “Life Cycle of Translational Research for Medical Interventions”. *Science* 321 (5894): 1298–1299.

Coughlin, Steven S./Catharie C. Nass/Linda W. Pickle/Bruce Trock/Greta Bunin. 1991. “Regression Methods for Estimating Attributable Risk in Population-based Case-Control Studies: A Comparison of Additive and Multiplicative Models”. *American Journal of Epidemiology* 133 (3): 305–313.

Cox, Christopher/Xiuhong Li. 2012. “Model-Based Estimation of the Attributable Risk: A Loglinear Approach.” *Computational statistics & data analysis* 56 (12): 4180–4189.

Cox, D.R. 1958. *The planning of experiments.* New York: John Wiley & Sons, Ltd.

Cox, Louis Anthony. 2013. “Caveats for Causal Interpretations of Linear Regression Coefficients for Fine Particulate (PM2.5) Air Pollution Health Effects”. *Risk Analysis* 33 (12): 2111–2125.

Craig, Peter/Cyrus Cooper/David Gunnell/Sally Haw/Kenny Lawson/Sally Macintyre/David Ogilvie/Mark Petticrew/Barney Reeves/Matt Sutton/Simon Thompson. 2012. “Using natural experiments to evaluate population health interventions: new Medical Research Council guidance.” *Journal of epidemiology and community health* 66 (12): 1182–6.

Crowson, Cynthia S/Terry M Therneau/W Michael O’fallon. 2009. *Attributable Risk Estimation in Cohort Studies*. Tech. rep.

Crump, Richard K./V. Joseph Hotz/Guido W. Imbens/Oscar A. Mitnik. 2009. *Dealing with limited overlap in estimation of average treatment effects.*

Czajka, John L./Sharon M. Hirabayashi/Roderick J.A. Little/Donald B. Rubin. 1992. “Projecting from advance data using propensity modeling: An application to income and tax statistics”. *Journal of Business and Economic Statistics* 10 (2): 117–131.

Czitrom, Veronica. 1997. “6. Introduction to Passive Data Collection”. In *Statistical Case Studies for Industrial Process Improvement*, 63–69. Society for Industrial / Applied Mathematics.

D ’agostino, Ralph B. 1998. “Tutorial in Biostatistics Propensity Score Methods for Bias Reduction in the Comparison of a Treatment To a Non-Randomized Control Group”. *STATISTICS IN MEDICINE Statist. Med* 17:2265–2281.

Dahabreh, Issa J./Radley C. Sheldrick/Jessica K. Paulus/Mei Chung/Vasileia Varvarigou/Haseeb Jafri/Jeremy A. Rassen/Thomas A. Trikalinos/Georgios D. Kitsios. 2012. “Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes". *European Heart Journal* 33 (15): 1893–1901.

Dehejia, Rajeev H./Sadek Wahba. 2002. "Propensity score-matching methods for nonexperimental causal studies". *Review of Economics and Statistics* 84 (1): 151–161. arXiv: `arXiv:1011.1669v3`.

Dekkers, O. M./E. von Elm/A. Algra/J. A. Romijn/J. P. Vandenbroucke. 2010. "How to assess the external validity of therapeutic trials: A conceptual approach". *International Journal of Epidemiology* 39 (1): 89–94.

Deubner, David C./William E. Wilkinson/Michael J. Helms/Herman A. Tyroler/ Curtis G. Hames. 1980. "LOGISTIC MODEL ESTIMATION OF DEATH ATTRIBUTABLE TO RISK FACTORS FOR CARDIOVASCULAR DISEASE IN EVANS COUNTY, GEORGIA". *American Journal of Epidemiology* 112 (1): 135–143.

Deza, Michel Marie/Elena Deza. 2009. "Encyclopedia of distances". In *Encyclopedia of Distances*, 1–590. Berlin, Heidelberg: Springer Berlin Heidelberg. arXiv: `0505065 [arXiv:gr-qc]`.

Dieng, Adji B/Dustin Tran/Rajesh Ranganath/John Paisley/David M Blei. 2016. *Variational Inference via χ Upper Bound Minimization.* Tech. rep. arXiv: `1611.00328v4`.

Dijkers, Marcel. 2013. "An e-newsletter from the Center on Knowledge Translation for Disability and Rehabilitation Research Introducing GRADE: a systematic approach to rating evidence in systematic reviews and to guideline development". *Presented in KT Update* 1 (5).

Djulbegovic, Benjamin/Gordon H. Guyatt. 2017a. "Progress in evidence-based medicine: a quarter century on". *The Lancet* 390 (10092): 415–423.

— . 2017b. "Progress in evidence-based medicine: a quarter century on". *The Lancet* 390 (10092): 415–423.

Djulbegovic, Benjamin/Gordon H. Guyatt/Richard E. Ashcroft. 2009. "Epistemologic inquiries in evidence-based medicine". *Cancer Control* 16 (2): 158–168.

Djulbegovic, Benjamin/Gordon Guyatt. 1976. "Evidence-based Medicine and the Theory of Knowledge". Chap. 3 in *Users Guide to Medical Literature*, 3rd ed. McGraw-Hill Education.

Doucet, Arnaud/Nando Freitas/Neil Gordon. 2001. "An Introduction to Sequential Monte Carlo Methods". In *Sequential Monte Carlo Methods in Practice*, 3–14. New York, NY: Springer New York.

DuMouchel, William/Daryl Pregibon. 2001. "Empirical bayes screening for multi-item associations". In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*, 67–76. New York, New York, USA: ACM Press.

Duan, N/Y Wang. 2012. "Heterogeneity of treatment effects". *Shanghai Arch Psychiatry* 24 (1): 50–51.

Dumouchel, William. 1999. “Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System”. *The American Statistician* 53 (3): 177–190.

Efron, Bradley/Trevor Hastie. 2016. *Computer age statistical inference: Algorithms, evidence, and data science*, 1–475.

Elwood, J. Mark. 1988. *Causal relationships in medicine : a practical system for critical appraisal.* 332. Oxford University Press.

Faxvaag, Arild/Trond S. Johansen/Vigdis Heimly/Line Melby/Anders Grimsmo. 2011. “Healthcare professionals’ experiences with EHR-system access control mechanisms”. *Studies in Health Technology and Informatics* 169:601–605.

Fisher, R A. 1935. *The design of experiments.* Oxford, England: Oxford & Boyd.

Fong, Christian/Chad Hazlett/Kosuke Imai. 2018. “Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements”. *The Annals of Applied Statistics* 12 (1): 156–177.

Frangakis, C. E./D. B. Rubin. 2002. “Principal stratification in causal inference”. *Biometrics* 58 (1): 21–29.

Franklin, Jessica M/Sara Dejene/Krista F Huybrechts/Shirley V Wang/Martin Kulldorff/Kenneth J Rothman. 2017. “A Bias in the Evaluation of Bias Comparing Randomized Trials with Nonexperimental Studies.” *Epidemiologic methods* 6 (1).

Franklin, Jessica M./Robert J. Glynn/David Martin/Sebastian Schneeweiss. 2019. “Evaluating the Use of Nonrandomized Real-World Data Analyses for Regulatory Decision Making”. *Clinical Pharmacology & Therapeutics*.

Franklin, Jessica M./Sebastian Schneeweiss. 2017. “When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials?” *Clinical Pharmacology & Therapeutics* 102 (6): 924–933.

Fredriksson, Peter/Per Johansson. 2008. “Dynamic Treatment Assignment”. *Journal of Business & Economic Statistics* 26 (4): 435–445.

Furler, John/Parker Magin/Marie Pirotta/Mieke van Driel. 2012. “Participant demographics reported in ”Table 1 of randomised controlled trials: a case of ”inverse evidence”?” *International Journal for Equity in Health* 11 (1): 14.

Gabler, Nicole B./Naihua Duan/Diana Liao/Joann G. Elmore/Theodore G. Ganiats/ Richard L. Kravitz. 2009. “Dealing with heterogeneity of treatment effects: Is the literature up to the challenge?” *Trials* 10:43.

Garfinkel, I/CF Manski/C Michalopoulos. 1992. “Micro experiments and macro effects”. In *Evaluating Welfare and Training Programs*. Cambridge, MA: Harvard University Press.

Gelman/J Hill. 2006. “Causal inference using regression on the treatment variable”. Chap. 9 in *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 167–198. Cambridge University Press.

Glazerman, Steven/Dan M. Levy/David Myers. 2003. “Nonexperimental versus Experimental Estimates of Earnings Impacts”. *Annals of the American Academy of Political and Social Science* 589 (1): 63–93.

Glynn, Adam N./Kevin M. Quinn. 2009. “An introduction to the augmented inverse propensity weighted estimator”. *Political Analysis* 18 (1): 36–56.

Glynn, Robert J./Joshua J. Gagne/Sebastian Schneeweiss. 2012. “Role of disease risk scores in comparative effectiveness research with emerging therapies”. *Pharmacoepidemiology and Drug Safety* 21 (SUPPL.2): 138–147. arXiv: `NIHMS150003`.

Glynn, Robert J./Sebastian Schneeweiss/Til Stürmer. 2006. “Indications for propensity scores and review of their use in pharmacoepidemiology”. *Basic and Clinical Pharmacology and Toxicology* 98 (3): 253–259. arXiv: `NIHMS150003`.

Good, I. J. 1959. *A Causal Calculus I.*

Good, I. J./Patrick Suppes. 1972. *A Probabilistic Theory of Causality.*, 67:245. 337. Amsterdam: North-Holland Publishing Company.

Goodman, N. 1947. “The Problem of Counterfactual Conditions”. *The Journal of Philosophy* 44 (5): 113–128.

Goshtasby, A. Ardeshir. 2012. “Similarity and Dissimilarity Measures”. In *Image Registration*, 7–66. London: Springer-Verlag.

Greenberg, B. G. 1953. “The use of analysis of convariance and balancing in analytical surveys.” *American journal of public health* 43 (6 :1): 692–699.

Greenland, S/J A Schwartzbaum/W D Finkle. 2000. “Problems due to small samples and sparse data in conditional logistic regression analysis.” *American journal of epidemiology* 151 (5): 531–9.

Greenland, Sander/Karsten Drescher. 1993. “Maximum Likelihood Estimation of the Attributable Fraction from Logistic Models”. *Biometrics* 49:865–872.

Greenland, Sander/Hal Morgenstern. 2001. “Confounding in Health Research”. *Annual Review of Public Health* 22 (1): 189–212.

Greenland, Sander/Judea Pearl/James M. Robins. 1999. “Causal Diagrams for Epidemiologic Research”. *Epidemiology* 10 (1): 37–48.

Greenland, Sander/James M Robins. 1988. *Reviews and Commentary CONCEPTUAL PROBLEMS IN THE DEFINITION AND INTERPRETATION OF ATTRIBUTABLE FRACTIONS*. Tech. rep.

Gretton, Arthur/Alex Smola/Jiayuan Huang/Marcel Schmittfull/Karsten Borgwardt/ Bernhard Schölkopf. 2009. *Covariate Shift by Kernel Mean Matching*. Tech. rep.

Grove, Charles C./R. A. Fisher. 1930. *Statistical Methods for Research Workers.*, 37:547. 10. Edinburgh: Oliver / Boyd. arXiv: `0-05-002170-2`.

Gu, X/PR Rosenbaum. 1993. “Comparison of multivariate matching methods: structures, distances, and algorithms”. *Journal of Computational and Graphical Statistics* 2:405–420.

Gulrajani, Ishaan/Faruk Ahmed/Martin Arjovsky/Vincent Dumoulin/Aaron Courville. *Improved Training of Wasserstein GANs Montreal Institute for Learning Algorithms*. Tech. rep.

Gupta, Somesh/Bhushan Kumar. *Sexually transmitted infections*. 1421.

Hahn, P. Richard/Jared Murray/Carlos M. Carvalho. 2017. "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects". *Ssrn*. arXiv: `1706.09523`.

Hainmueller, Jens. 2011. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies". *Political Analysis* 16:25–46.

Halpern, Elkan F. 2014. "Behind the Numbers: Inverse Probability Weighting". *Radiology* 271 (3): 625–628.

Hamming, R. 1950. "Tech. J. 29". *Bell Syst* 147.

Hansen, BB. 2008. "The prognostic analogue of the propensity score". *Biometrika* 95 (2): 481–488.

Hansen, Ben B. 2004. "Full matching in an observational study of coaching for the SAT". *Journal of the American Statistical Association* 99 (467): 609–618.

Hansen, Ben B./Stephanie Olsen Klopfer. 2006. "Optimal full matching and related designs via network flows". *Journal of Computational and Graphical Statistics* 15 (3): 609–627.

Harbour, R/J Miller. 2001. “A new system for grading recommendations in evidence based guidelines.” *BMJ (Clinical research ed.)* 323 (7308): 334–6.

Harpaz, Rave/Santiago Vilar/William DuMouchel/Hojjat Salmasian/Krystl Haerian/ Nigam H Shah/Herbert S Chase/Carol Friedman. 2013. “Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions”. *Journal of the American Medical Informatics Association* 20 (3): 413–419.

Hartley, Paul/Yue Shentu/Patricia Betz-Schiff/Gregory T. Golm/Christine McCrary Sisk/Samuel S. Engel/R. Ravi Shankar. 2015. “Efficacy and Tolerability of Sitagliptin Compared with Glimepiride in Elderly Patients with Type 2 Diabetes Mellitus and Inadequate Glycemic Control: A Randomized, Double-Blind, Non-Inferiority Trial”. *Drugs & Aging* 32 (6): 469–476.

Hausman, Daniel Murray. 2005. “Causal Relata: Tokens, Types, or Variables? on JSTOR”. *Erkenntis* 63:33–54.

Hazlett, Chad. 2016. “Kernel Balancing: A Flexible Non-Parametric Weighting Procedure for Estimating Causal Effects”. *Ssrn.* arXiv: 1605.00155.

Heckman, James/Hidehiko Ichimura/Jeffrey Smith/Petra Todd. 1998. “Characterizing Selection Bias Using Experimental Data”. *Econometrica* 66 (5).

Heinze, Oliver/Markus Birkle/Lennart Köster/Björn Bergh. 2011. “Architecture of a consent management suite and integration into IHE-based regional health information networks”. *BMC Medical Informatics and Decision Making* 11 (1): 58.

Hellerstein, J. K./Guido W Imbens. 1999. “Imposing moment restrictions from auxiliary data by weighting”. *Review of Economics and Statistics* 81:1–14.

Hemkens, Lars G/Despina G Contopoulos-Ioannidis/John P A Ioannidis. 2016. “Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey.” *BMJ (Clinical research ed.)* 352:i493.

Henrion, Max. 1987. “Practical issues in constructing a Bayes’ belief networks”. In *Uncertainty in artificial intelligence*, 132–199. New York: North-Holland.

Hernán, Miguel A/James M Robins. 2006. “Estimating causal effects from epidemiological data”. *J Epidemiol Community Health* 60:578–586.

— . 2016. “Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available.” *American journal of epidemiology* 183 (8): 758–64.

Higgins, Julian P T/Douglas G Altman/Peter C Gøtzsche/Peter Jüni/David Moher/Andrew D Oxman/Jelena Savović/Kenneth F Schulz/Laura Weeks/Jonathan A C Sterne. 2011. “The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials”. *BMJ* 343.

Hill, A B. 1965. “THE ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION?” *Proceedings of the Royal Society of Medicine* 58 (5): 295–300.

Hill, Jennifer L. 2011. “Bayesian Nonparametric Modeling for Causal Inference”. *Journal of Computational and Graphical Statistics* 20 (1): 217–240.

Hirano, Kieisuke/Guido W Imbens/Geert Ridder '. 2003. "EFFICIENT ESTIMATION OF AVERAGE TREATMENT EFFECTS USING THE ESTIMATED PROPENSITY SCORE". *Econometrica* 71 (4): 1161–1189.

Ho, Daniel E./Kosuke Imai/Gary King/Elizabeth A. Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference". *Political Analysis* 15 (3): 199–236.

Höfler, M. 2005. "Causal inference based on counterfactuals". *BMC Medical Research Methodology* 5 (1): 28.

Holland, Paul W. 1988. "Causal Inference, Path Analysis and Recursive Structural Equations Models". *Sociological Methodology* 18:449–484.

— . 1986. "Statistics and Causal Inference". *Journal of the American Statistical Association* 81 (396): 945–960.

Holland, Paul W./Donald B. Rubin. 1987. "CAUSAL INFERENCE IN RETROSPECTIVE STUDIES". *ETS Research Report Series* 1987 (1): 203–231.

Hripcsak, George/David J Albers. 2012. "Next-generation phenotyping of electronic health records". *JAMIA* 20:117–121.

Hripcsak, George/Jon D Duke/Nigam H Shah/Christian G Reich/Vojtech Huser/Martijn J Schuemie/Marc A Suchard/Rae Woong Park/Ian Chi Kei Wong/Peter R Rijnbeek/Johan van der Lei/Nicole Pratt/G Niklas Norén/Yu-Chuan Li/Paul E Stang/David Madigan/Patrick B Ryan. 2015. "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers." *Studies in health technology and informatics* 216:574–8.

Hsu, Jesse Y./José R. Zubizarreta/Dylan S. Small/Paul R. Rosenbaum. 2015. “Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods”. *Biometrika* 102 (4): 767–782.

Hyman, Ray. 1982. *Quasi-Experimentation: Design and Analysis Issues for Field Settings (Book).*

Iacus, Stefano M./Gary King/Giuseppe Porro. 2012. “Causal inference without balance checking: Coarsened exact matching”. *Political Analysis* 20 (1): 1–24.

Ihrie, John. 2019. *EBGM Disproportionality Scores for Adverse Event Data Mining.*

Illari, Phyllis McKay/Federica Russo/Jon Williamson. 2011. “Causality, theories and medicine”. In *Causality in the Sciences.* Oxford.

Imai, Kosuke. 2013. *Matching and Weighting Methods for Causal Inference.*

Imai, Kosuke/Gary King/Clayton Nall. 2009. “The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation”. *Statistical Science* 24 (1): 29–53.

Imai, Kosuke/Gary King/Elizabeth Stuart. 2008. “Misunderstandings Among Experimentalists and Observationalists about Causal Inference”. *Journal of the Royal Statistical Society, Series A* 171, part.

Imai, Kosuke/Marc Ratkovic. 2013. “Covariate balancing propensity score”. *J. R. Statist. Soc. B* 76 (1): 243–263.

Imbens/Rubin. 2009. “Chapter 15 Design in Observational Studies: Matching to Ensure Balance in Covariate Distributions”. In *Causal Inference Part II.*

Imbens, G/G King/D McKenzie/G Ridder. 2009. "On the Benefits of Stratification in Randomized Experiments". *unpublished manuscript, Department of Economics, Harvard University.*

Imbens, Guido W. 2009. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua". *unpublished manuscript, Department of Economics, Harvard University.*

— . 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review NONPARAMETRIC ESTIMATION OF AVERAGE TREATMENT EFFECTS UNDER EXOGENEITY: A REVIEW*". *Source: The Review of Economics and Statistics* 86 (1): 4–29.

Imbens, Guido W./Donald B. Rubin. 2015. *Causal inference: For statistics, social, and biomedical sciences an introduction*, 1–625. Cambridge University Press. arXiv: `arXiv:1011.1669v3`.

Ioannidis, John P A. 2014. "How to make more published research true." *PLoS medicine* 11 (10): e1001747.

Ishii, H. 1972. "Symposium I. Hormone therapy in otolaryngology, especially short term general dosage of adrenal cortex hormones (Japanese)". *Journal of Otolaryngology of Japan* 75 (12): 1464–1465.

Jager, K.J./C. Zoccali/A. MacLeod/F.W. Dekker. 2008. "Confounding: What it is and how to deal with it". *Kidney International* 73 (3): 256–260.

Jamerson, Kenneth/Michael A. Weber/George L. Bakris/Björn Dahlöf/Bertram Pitt/ Victor Shi/Allen Hester/Jitendra Gupte/Marjorie Gatlin/Eric J. Velazquez. 2008.

"Benazepril plus Amlodipine or Hydrochlorothiazide for Hypertension in High-Risk Patients". *New England Journal of Medicine* 359 (23): 2417–2428.

Johansson, Fredrik D./Uri Shalit/David Sontag. 2016. "Learning Representations for Counterfactual Inference". arXiv: `1605.03661`.

Kallus, Nathan. 2017. "A Framework for Optimal Matching for Causal Inference". In *Artificial Intelligence and Statistics (AISTATS)*, 372–381. arXiv: `1606.05188`.

— . 2016. "Causal inference by minimizing the dual norm of bias: Kernel matching & weighting estimators for causal effects". *CEUR Workshop Proceedings* 1792:18–28. arXiv: `1606.05188`.

— . 2018a. "DeepMatch: Balancing Deep Covariate Representations for Causal Inference Using Adversarial Training". arXiv: `1802.05664`.

— . 2018b. "Optimal a priori balance in the design of controlled experiments". *J. R. Statist. Soc. B* 80 (1): 85–112. arXiv: `1312.0531v4`.

Kang, Joseph D Y/Joseph L Schafer. 2007. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data". *Statistical Science* 22 (4): 523–539. arXiv: `0804.2958v1`.

Kaplan, David. 2018. "Causal Inference for Observational Studies". *The Journal of Infectious Diseases.*

Karanis, Y.B./F.A. Bermudez Canta/L. Mitrofan/H. Mistry/C. Anger. 2016. "'Research' vs 'real world' patients: the representativeness of clinical trial participants". *Annals of Oncology* 27 (suppl_6).

Kaur, Deepinder. 2014. “A Comparative Study of Various Distance Measures for Software fault prediction”. *International Journal of Computer Trends and Technology* 17 (3).

Keele, Luke/JR Zubizarreta. 2014. “Optimal Multilevel Matching in Clustered Observational Studies : A Case Study of the School Voucher System in Chile”. *arXiv preprint arXiv:1409.8597*: 1–37. arXiv: `1409.8597`.

Kelcey, B. 2013. “Propensity Score Matching Within Prognostic Strata”. In *SREE 2013 Conference*.

Kelsey, Jl/WD Thompson. 1986. *Methods in observational epidemiology.*, 432. Oxford University Press.

Kempthorne, O. 1955. “The randomizaton theory of experimental inference”. *Journal of American Statistics* 50:946–967.

Kerbyson, Darren J./Kevin J. Barker/Abhinav Vishnu/Adolfy Hoisie. 2014. “A performance comparison of current HPC systems: Blue Gene/Q, Cray XE6 and InfiniBand systems”. *Future Generation Computer Systems* 30 (1): 291–304. arXiv: `1210.4852`.

Kernan, Walter N/Catherine M Viscoli/Robert W Makuch/Lawrence M Brass/Ralph I Horwitz. 1999. “Stratified Randomization for Clinical Trials”. *J Clin Epidemiol* 521 (52): 19–2619.

Kim, Hae-Young. 2017. “Statistical notes for clinical researchers: Risk difference, risk ratio, and odds ratio”. *Restorative Dentistry & Endodontics* 42 (1): 72.

Kim, Jin H/Judea Pearl. *A COMPUTATIONAL MODEL FOR CAUSAL AND DIAGNOSTIC REASONING IN INFERENCE SYSTEMS*. Tech. rep.

King. 2011a. “Matching Methods for Causal Inference with”. *Matching Methods for Casual Inference*: 27.

King, Gary. 2011b. “Comparative Effectiveness of Matching Methods for Causal Inference *”. *Director of Health Research and Outcomes*.

King, Gary/Richard Nielsen. 2018. “Why Propensity Scores Should Not Be Used for Matching — GARY KING”. *Internal report, Harvard University.* 3 (10).

Kish, L. 1965. “Survey Sampling”. Chap. Chapter 14 in *Survey Sampling*. New York: Wiley.

Kleinbaum, David G./Lawrence L. Kupper/Hal Morgenstern. 1982. *Epidemiologic research : principles and quantitative methods*. 529. John Wiley & Sons.

Kleinberg, Samantha/George Hripcsak. 2011. “A review of causal inference for biomedical informatics”. *Journal of Biomedical Informatics* 44 (6): 1102–1112.

Kohonen, Teuvo./M. R. Schroeder/T. S. Huang. 2001. *Self-organizing maps*. 501. Berlin: Springer-Verlag. arXiv: `arXiv:1011.1669v3`.

Kompan, Michal. 2011. *Information retrieval*. Butterworth-Heinemann.

Kooperberg, Charles/Diana B Petitti. 1991. *Using Logistic Regression to Estimate the Adjusted Attributable Risk of Low Birthweight in an Unmatched Case-Control Study*. Tech. rep. 5.

Kravitz, Richard L RL/Naihua Duan/Joel Braslow. 2004. “Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages”. *Milbank Q* 82 (4): 661–687.

Kvart, Igal. 2004. “Causation: Probabilistic and Counterfactual Analysis”. In *Collins, Hall and Paul*, 359–387.

Lachin, JM/JP Matis/LJ Wei. 1988. “Randomizations in clinical trails, conclusions and recommendations”. *Control Clin Trails.* 9:365–374.

Laffers, Lukas/Giovanni Mellace. 2016. “Identification of the Average Treatment Effect when SUTVA is violated”. *Working Paper*: 1–30.

Lalonde, Robert J. 1986. “Evaluating the Econometric Evaluations of Training Programs with Experimental Data”. *The American Economic Review* 76 (4): 604–620.

Lance, G. N./W. T. Williams. 1966. “Computer Programs for Hierarchical Polythetic Classification (”Similarity Analyses”)”. *The Computer Journal* 9 (1): 60–64.

Lance, GN/WT Williams. 1967. “Hierarchical Systems”. In *A general theory of classificatory sorting strategies.*

Langevin, Helene M./Elisa E. Konofagou/Gary J. Badger/David L. Churchill/James R. Fox/Jonathan Ophir/Brian S. Garra. 2004. “Tissue displacements during acupuncture using ultrasound elastography techniques”. *Ultrasound in Medicine and Biology* 30 (9): 1173–1183. arXiv: 1504.01132.

Lechner, M. 2001. “Identification and Estimation of Causal Effects of Multiple Treat- ments under the Conditional Independence Assumption”. In *Econometric Evaluation of Labour Market Policies*, ed. by M Lechner/D Pfeiffer, 43–58. Heidelberg: Physica.

Lee, BK/J Lessler/EA Stuart. 2010. “Improving Propensity score weighting using machine learning”. *Statistics in Medicine* 29:337–346.

Leger, A S S. 1994. *Statistical Models in Epidemiology*, 48:607–607. 6. Oxford University Press.

Levin, M L. 1953. “The occurrence of lung cancer in man.” *Acta - Unio Internationalis Contra Cancrum* 9 (3): 531–41.

Leviton, Alan. 1973. “DEFINITIONS OF ATTRIBUTABLE RISK”. *American Journal of Epidemiology* 98 (3): 231–231.

Lewis, D. 1973a. “Causation”. *Journal of Philosophy* 70:556–567.

— . 1973b. *Counterfactuals*. Oxford: Blackwell.

Li, Fan/Kari Lock Morgan/Alan M. Zaslavsky. 2018. “Balancing Covariates via Propensity Score Weighting”. *Journal of the American Statistical Association* 113 (521): 390–400. arXiv: 1404.1785.

Linden, A/PR Yarnold. 2017. “Using classification tree analysis to generate propensity score weights”. *J Eval Clin Pract* 23 (4): 703–712.

Liu, Z/Q Fang/J Zuo/V Minhas/C Wood/T Zhang. 2018. “The world-wide incidence of Kaposi’s sarcoma in the HIV / AIDS era”. *HIV Medicine* 19 (5): 355–364.

Longford, N T. 1999. “Selection bias and treatment heterogeneity in clinical trials.” *Statistics in medicine* 18 (12): 1467–74.

Lourenco, F./V. Lobo/F. Bacao. 2004. “Binary-based similarity measures for categorical data and their application in Self-Organizing Maps”. *Citeseer*: 1–18.

Lunceford, Jared K. JK/Marie Davidian. 2004. “Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study”. *Stat Med.* 23 (19): 2937–2960.

Lunt, Mark. 2014. “Selecting an Appropriate Caliper Can Be Essential for Achieving Good Balance With Propensity Score Matching”. *American Journal of Epidemiology* 179 (2): 226–235.

Lyon, A. 1967. “Causality”. *British Journal for the Philosophy of Science* 18:1–20.

Mackie, J L. 1965. “Causes and Conditions”. *American Philosophical Quarterly* 12:245–265.

Mackie, J. L. 1980. *The Cement of the Universe.* 1st ed. Oxford University Press. arXiv: `arXiv:1011.1669v3`.

Markush, Robert E. 1977. “LEVIN’S ATTRIBUTABLE RISK STATISTIC FOR ANALYTIC STUDIES AND VITAL STATISTICS”. *American Journal of Epidemiology* 105 (5): 401–406.

McCaffrey, DF/G Ridgeway/Morral AR/AR Morral. 2004a. "Propensity Score Estimation with Boosted Regression for evaluating causal effects in observational studies". *Psychological Methods* 9:403–425.

McCaffrey, Daniel F./Greg Ridgeway/Andrew R. Morral. 2004b. "Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods* 9 (4): 403–425.

McNamee, R. 2005. "Regression modelling and other methods to control confounding." *Occupational and environmental medicine* 62 (7): 500–6, 472.

Mense, Alexander/Franz Hoheiser-Pförtner/Martin Schmid/Harald Wahl. 2013. "Concepts for a standard based cross-organisational information security management system in the context of a nationwide EHR." *Studies in health technology and informatics* 192:548–52.

Mescheder, Lars/Andreas Geiger/Sebastian Nowozin. 2018. *Which Training Methods for GANs do actually Converge?* Tech. rep. arXiv: 1801.04406v4.

Meyer, BD. 1995. "Natural and quasi-experiments in economics". *Journal of Business and Economic Studies* 13:151–161.

Miettinen, Olli S. 1974. "PROPORTION OF DISEASE CAUSED OR PREVENTED BY A GIVEN EXPOSURE, TRAIT OR INTERVENTION". *AMERICAN JOURNAL or EPIDEMIOLOGY* 99 (5).

Ming, Kewei/Paul R. Rosenbaum. 2001. *A Note on Optimal Matching with Variable Controls Using the Assignment Algorithm.*

Moher, D/A R Jadad/P Tugwell. 1996. “Assessing the quality of randomized controlled trials. Current issues and future directions.” *International journal of technology assessment in health care* 12 (2): 195–208.

Moher, David/Sally Hopewell/Kenneth F Schulz/Victor Montori/Peter C Gøtzsche/ P J Devereaux/Diana Elbourne/Matthias Egger/Douglas G Altman. 2010. “CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials”. *BMJ* 340.

Morgan, Stephen L/David J Harding. 2006. “Matching Estimators of Causal Effects Prospects and Pitfalls in Theory and Practice”. *Sociological Methods & Research Sage Publications* 35 (10).

Morgan, Stephen L./Jennifer J. Todd. 2008. “6. A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects”. *Sociological Methodology* 38 (1): 231–282.

Morgan, Stephen L/Christopher Winship. 2007. *Counterfactuals and causal inference : methods and principles for social research / Stephen L. Morgan, Christopher Winship.* 2nd ed. Accessed from http://nla.gov.au/nla.cat-vn4231122. Cambridge: Cambridge University Press.

Morris, Genevieve/Audacious Inquiry/Greg Farnum/Scott Afzal. 2014. *Patient Identification and Matching Final Report.*

Muller, Mervin E. 1966. *Review: J. M. Hammersley, D. C. Handscomb, Monte Carlo Methods ; Yu. A. Shreider, Methods of Statistical Testing/Monte Carlo Method*, 37:532–538. 2. Springer Netherlands.

Murdoch, TB/AS Detsky. 2013. "The inevitable application of big data to health care". *JAMA* 309 (13): 1352–1352.

Neyman, J. 1923. "On the Application of Probability Theory of Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5 (4): 465–472.

Neyman, J./K. Iwaszkiewicz/St. Kolodziejczyk. 1935. "Statistical Problems in Agricultural Experimentation". *Supplement to the Journal of the Royal Statistical Society* 2 (2): 107.

Nielsen, Richard A. 2016. "Case Selection via Matching". *Sociological Methods and Research* 45 (3): 569–597.

Nowozin, Sebastian/Botond Cseke/Ryota Tomioka. 2016. *f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization.* arXiv: `1606.00709`.

Open Science Collaboration. 2015. "Estimating the reproducibility of psychological science". *Science* 349 (6251): aac4716.

Papanicolas, I./P.C. Smith. 2014. *Encyclopedia of Health Economics*, 386–394. Elsevier.

Parsons, L. 2001. "Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching..." In *The 26th Annual SAS Users Group International Conference*, 214–226.

Payne, RW. 2015. *The Design and Analysis of Experiments*, ed. by Wiley, 107:772–785. New York.

Pearl, J. 2000. *Causality.* Cambridge, England: Cambridge University Press.

Pearl, Judea. 2010. “3. The Foundations of Causal Inference”. *Sociological Methodology* 40 (1): 75–149.

— . 1993. “[Bayesian Analysis in Expert Systems]: Comment: Graphical Models, Causality and Intervention”. *Statistical Science* 8 (3): 266–269.

— . 1995. “Causal Diagrams for Empirical Research”. *Biometrika* 82 (4): 669–688.

— . 2011. “Principal stratification–a goal or a tool?” *The international journal of biostatistics* 7 (1): 20.

— . 2012. “The Do-Calculus Revisited”. arXiv: 1210.4852.

Peters, Jonas/Joris Mooij/Dominik Janzing/Bernhard Schölkopf. 2013. “Causal Discovery with Continuous Additive Noise Models”. arXiv: 1309.6779.

Pourhoseingholi, Mohamad Amin/Ahmad Reza Baghestani/Mohsen Vahedi. 2012. “How to control confounding effects by statistical analysis.” *Gastroenterology and hepatology from bed to bench* 5 (2): 79–83.

Pratt, John W/Robert Schlaifer. 1984. “On the nature and discovery of structure”. *Journal of the American Statistical Association* 79 (9-21): 29–33.

Ratkovic, Marc. 2014. “Balancing within the margin: Causal effect estimation with support vector machines”. *Department of Politics, Princeton University, Princeton, NJ.*

Ray, Greg. 1992. “Probabilistic causality reexamined”. Chap. Chapter 6: in *Erkenntnis*, ed. by David Bourget/David Chalmers, 36:219–244. 2. New York: Cambridge University Press.

Reuter, Peter. 1991. “On the consequences of toughness”. *Biometrika Biometrika Trust Biometrika Biometrika Trust Advance Access* 95 (4): 55–80.

Robins, J M/M A Hernán/B Brumback. 2000. “Marginal structural models and causal inference in epidemiology.” *Epidemiology (Cambridge, Mass.)* 11 (5): 550–60.

Robins, JM/S Greenland. 1994. “Adjusting for differential rates of prophylaxis therapy for PCP in high versus low dose AZT treatment arms in an AIDS randomized trial”. *J Am Stat* 89:737–749.

Robins, James M./Andrea Rotnitzky/Lue Ping Zhao. 1994. “Estimation of Regression Coefficients When Some Regressors Are Not Always Observed”. *Journal of the American Statistical Association* 89 (427): 846.

Robins, James. 1986. “A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect”. *Mathematical Modelling* 7 (9-12): 1393–1512.

Rockhill, B/B Newman/C Weinberg. 1998. “Use and misuse of population attributable fractions.” *American journal of public health* 88 (1): 15–9.

Rosen, Deborah A. 1978. *In Defense of a Probabilistic Theory of Causality.*

Rosenbaum, P. R./D. B. Rubin. 1983a. *Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome.*

Rosenbaum, P R/D B Rubin. 1985a. “The bias due to incomplete matching.” *Biometrics* 41 (1): 103–16.

Rosenbaum, PR Paul R./Donald B. DB Rubin. 1985b. “Constructing a control group using multivariate matched sampling methods that incorporate the propensity score”. *The American Statistician* 39 (1): 33–38.

Rosenbaum, PR/Richard N Ross/Jeffrey H Silber. 2007. “Minimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatment for Ovarian Cancer”. *Journal of American Statistical Association* 102 (477).

Rosenbaum, Paul R. 1991. “A Characterization of Optimal Designs for Observational Studies”. *Source Journal of the Royal Statistical Society. Series B (Methodological)* 53 (3): 597–610.

— . 1989. “Optimal Matching for Observational Studies”. *Source Journal of the American Statistical Association* 8411888 (408): 1024–1032.

Rosenbaum, Paul R. 2012. “Optimal Matching of an Optimally Chosen Subset in Observational Studies”. *Journal of Computational and Graphical Statistics* 21 (1): 57–71.

Rosenbaum, Paul R./Donald B. Rubin. 1984. “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score”. *Journal of the American Statistical Association* 79 (387): 516.

Rosenbaum, Paul R/Donald B Rubin. 1983b. “The central role of the propensity score in observational studies for causal effects”. *Biometrika* 70 (1): 41–55.

Rosenblatt, Murray. 1956. “Remarks on Some Nonparametric Estimates of a Density Function”. *The Annals of Mathematical Statistics* 27 (3): 832–837.

Rothman, K J. 2000. “Declaration of Helsinki should be strengthened.” *BMJ (Clinical research ed.)* 321 (7258): 442–5.

Rothman, Kenneth J. 1976. “Causes”. *American Journal of Epidemiology* 104 (6): 587–592.

Rothman, Kenneth J./Sander Greenland/Timothy L. Lash. 1998. *Modern Epidemiology - Rothman.* 1–758. Wolters Kluwer Health/Lippincott Williams & Wilkins.

Rothwell, Peter M. 2006. “Factors That Can Affect t he External Validity of Randomized Controlled Trials”. *PLoS Clinical Trials* 1 (1): e9.

Rubin, By Donald B. 2012. *Matched Sampling for Causal E ff ects.* 771–787. June. Cambridge University Press.

Rubin, D. B. 1980. “Randomization analysis of experimental data: The Fisher randomization test comment”. *Journal of American Statistical Association* 75 (371): 591–593.

Rubin, DB. 1990. “Comment: Neyman (1923) and causal inference in experiments and observational studies”. *Statistical Science* 5:472–480.

— . 1985. “The use of propensity scores in applied Bayesian inference”. *Bayesian Statistics* 2:463–472.

— . 2001. “Using propensity scores to help design observational studies: application to the tobacco litigation”. *Health Services & Outcomes Research Methodology* 2:169–188.

Rubin, DB/N Thomas. 2000. “Combining propensity score matching with additional adjustment for prognostic covariates”. *Journal of the American Statistical Association* 95:573–585.

Rubin, Donald B. DB. 2007. “The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials”. *Statistics in Medicine* 26 (1): 20–36.

Rubin, Donald B. DC/Neal Thomas. 1992. “Characterising the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions”. *Biometrika* 79 (4): 797–809.

Rubin, Donald B. 1977. “Assignment to Treatment Group on the Basis of a Covariate”. *Journal of Educational Statistics* 2 (1): 1.

— . 1978. “Bayesian Inference for Causal Effects: The Role of Randomization”. *The Annals of Statistics* 6 (1): 34–58.

— . 2005. *Causal Inference Using Potential Outcomes: Design, Modeling, Decisions.*

— . 1974. “Estimating causal effects of treatments in randomized and nonrandomized studies.” *Journal of Educational Psychology* 66 (5): 688–701.

— . 1973a. “Matching to Remove Bias in Observational Studies”. *Biometrics* 29 (1): 159.

— . 1973b. “The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies”. *Biometrics* 29 (1): 185.

Rubin, Donald B. 1986. "Which Ifs Have Causal Answers". *Journal of the American Statistical Association* 81 (396): 961–962.

Sackett, D L/W M Rosenberg/J A Gray/R B Haynes/W S Richardson. 1996. "Evidence based medicine: what it is and what it isn't." *BMJ (Clinical research ed.)* 312 (7023): 71–2.

Sakaeda, Toshiyuki/Akiko Tamon/Kaori Kadoyama/Yasushi Okuno. 2013. "Data Mining of the Public Version of the FDA Adverse Event Reporting System". *International Journal of Medical Sciences* 10 (7): 796–803.

Salmon, W. 1988. "Probabilistic Causality". In *Causality and explanation*, 208–232. Oxford University Press.

Scharfstein, Daniel O./Andrea Rotnitzky/James M. Robins. 1999. "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models: Rejoinder". *Journal of the American Statistical Association* 94 (448): 1135.

Schlesselman, James J./Paul D. Stolley. 1982. *Case-control studies : design, conduct, analysis.* 354. Oxford University Press.

Schmidt, Carsten Oliver/Thomas Kohlmann. 2008. "When to use the odds ratio or the relative risk?" *International Journal of Public Health* 53 (3): 165–167.

Schochet, Peter Z. 2010. "Is regression adjustment supported by the Neyman model for causal inference?" *Journal of Statistical Planning and Inference* 140 (1): 246–259.

Schuemie, Martijn J/Marc A Suchard/Patrick Ryan. 2016. "Single studies using the CohortMethod package": 1–20.

Schulz, Kenneth F/Douglas G Altman/David Moher. 2010. “CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials”. *BMJ* 340:c332.

Schwab, Patrick/Lorenz Linhardt/Walter Karlen. 2018. *Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks.* arXiv: `1810.00656 [cs.LG]`.

Setoguchi, S/Sebastian Schneeweiss/M. Alan Brookhart. 2008. “Evaluating uses of data mining techniques in propensity score estimation: A simulation study”. *Pharmacoepidemiology and Drug Safety* 17:546–555.

Seung-Seok, Choi/Cha Sung-Hyuk/Charles C Tappert/Computer Science. 2014. “A Survey of Binary Similarity and Distance Measures.” *Journal of Systemics, Cybernetics & Informatics* 8 (1): 43–48.

Shadish, WR/TD Cook/DT Campbell. 2002. *Experimental and quasi-experimental designs for generalised causal inference.* Boston, MA: Houghton Mifflin Company.

Shah, N H. 2012. “Translational bioinformatics embraces big data.” *Yearbook of medical informatics* 7:130–4.

Shalit, Uri/Fredrik D Johansson/David Sontag. 2017. “Estimating individual treatment effect: generalization bounds and algorithms”. In *ICML*.

Sherman, Rachel E./Steven A. Anderson/Gerald J. Dal Pan/Gerry W. Gray/Thomas Gross/Nina L. Hunter/Lisa LaVange/Danica Marinac-Dabic/Peter W. Marks/Melissa A. Robb/Jeffrey Shuren/Robert Temple/Janet Woodcock/Lilly Q. Yue/

Robert M. Califf. 2016. “Real-World Evidence — What Is It and What Can It Tell Us?” *New England Journal of Medicine* 375 (23): 2293–2297.

Shirkhorshidi, Ali Seyed/Saeed Aghabozorgi/Teh Ying Wah. 2015. “A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data.” *PloS one* 10 (12): e0144059.

Sinclair, John C. 2003. “Weighing risks and benefits in treating the individual patient.” *Clinics in perinatology* 30 (2): 251–68.

Smith, Herbert L. 1992. “6 MATCHING WITH MULTIPLE CONTROLS TO ESTIMATE TREATMENT EFFECTS IN OBSERVATIONAL STUDIES”. *Sociological Methodology* 27:325–353.

Smith, Jeffrey A./Petra E. Todd. 2005. “Does matching overcome LaLonde’s critique of nonexperimental estimators?” *Journal of Econometrics* 125 (1-2 SPEC. ISS.): 305–353.

Sobel, M E. 2000. “Causal inference in the social sciences”. *Journal of the American Statistical Association* 95 (450): 647–651.

Sobel, M. 1994. “Causal inference in latent variable models.” In *atent variables analysis: Applications for developmental research*, 3–35. Thousand Oaks, CA: Sage Publications, Inc.

Sobel, Michael E. 1995. “Causal Inference in the Social and Behavioral Sciences”. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, 1–38. Boston, MA: Springer US.

Sorenson, Bent E. 2005. “Gra_Caus”: 1–4.

Spiel, Christiane/Dominik Lapka/Petra Gradinger/Eva Maria Zodlhofer/Ralph Reimann/Barbara Schober/Petra Wagner/Alexander von Eye. 2008. “A Euclidean Distance-Based Matching Procedure for Nonrandomized Comparison Studies”. *European Psychologist* 13 (3): 180–187.

Splawa-Neyman, Jerzy/D. M. Dabrowska/T. P. Speed. 1990. *On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.*

Srinivas, Sampath. 2013. “A Generalization of the Noisy-Or Model”. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence (UAI1993).*

Steckler, Allan/Kenneth R McLeroy. 2008. “The importance of external validity.” *American journal of public health* 98 (1): 9–10.

Steyer, Rolf/Christof Nachtigall/Olivia Wüthrich-Martone/Katrin Kraus. 2002. “Causal Regression Models III: Covariates, Conditional, and Unconditional Average Causal Effects 1”. *Methods of Psychological Research Online* 7 (1).

Stuart, Elizabeth A. 2010. “Matching methods for causal inference: A review and a look forward.” *Statistical science : a review journal of the Institute of Mathematical Statistics* 25 (1): 1–21.

Stuart, Elizabeth A/Catherine P Bradshaw/Philip J Leaf. 2015. “Assessing the generalizability of randomized trial results to target populations.” *Prevention science : the official journal of the Society for Prevention Research* 16 (3): 475–85.

Susser, M. 1973. *Causal thinking in the health sciences: concepts and strategies of epidemiology.* Oxford University Press Ind.

Tao, Chenyang/Liqun Chen/Ricardo Henao/Jianfeng Feng/Lawrence Carin Duke. 2018. *Chi-square Generative Adversarial Network.*

Tejero, Antonio/Isabel de la Torre. 2012. “Advances and Current State of the Security and Privacy in Electronic Health Records: Survey from a Social Perspective”. *Journal of Medical Systems* 36 (5): 3019–3027.

Thadhani, Ravi. 2006. *Formal trials versus observational studies.* Oxford PharmaGenesis.

Thoemmes, Felix/Anthony D. Ong. 2016. “A Primer on Inverse Probability of Treatment Weighting and Marginal Structural Models”. *Emerging Adulthood* 4 (1): 40–59.

Thrusfield, Michael. 2017. “Observational studies”. In *Veterinary Epidemiology: Fourth Edition*, 319–338. arXiv: `arXiv:1011.1669v3`.

VanderWeele, Tyler J/Miguel A Hernán. 2013. “Causal Inference Under Multiple Versions of Treatment.” *Journal of causal inference* 1 (1): 1–20.

Vanderweele, Tyler J. 2011. “Principal stratification–uses and limitations.” *The international journal of biostatistics* 7 (1).

Vansteelandt, S./R.M. Daniel. 2014. “On regression adjustment for the propensity score”. *Statistics in Medicine* 33 (23): 4053–4072.

Velasco, Eduardo. 2010. “Exlcusion Criteria”. Chap. Exclusion in *Encyclopedia of Research Design*, ed. by Neil Salkind. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc.

Vilar, Santiago/Rave Harpaz/Lourdes Santana/Eugenio Uriarte/Carol Friedman. 2012. “Enhancing adverse drug event detection in electronic health records using molecular structure similarity: application to pancreatitis.” *PloS one* 7 (7): e41471.

Vogt, WP/B Johnson. 2011. *Dictionary of statistics & methodology: a nontechnical guide for the social sciences.* 4th, xviii, 437 p. Washington DC: SAGE Publications Inc.

Wacholder, S/C R Weinberg. 1982. “Paired versus two-sample design for a clinical trial of treatments with dichotomous outcome: power considerations.” *Biometrics* 38 (3): 801–12.

Wager, Stefan/Susan Athey. 2018. “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”. *Journal of the American Statistical Association* 113 (523): 1228–1242. arXiv: `1510.04342`.

Wales, Jackie A. 2009. “Can treatment trial samples be representative?” *Behaviour Research and Therapy* 47 (10): 893–896.

Weiskopf, Nicole G/George Hripscak/Sushmita Swaminathan/Chunhua Weng. 2013. “Defining and measuring completeness of electronic health records for secondary use”. *Journal of Biomedical Informatics* 46:830–836.

Weitzen, S/KL Lapane/AY Toleando/AL Hume/V Mor. 2004. “Principles for modeling propensity scores in medical research: a systematic literature review”. *Pharmacoepidemiology and Drug Safety* 127:626–639.

Westreich, D. 2010. “Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression”. *Journal of Clinical Epidemiology* 63 (8): 826–833.

Wilks, SS. 1932. “On the distribution of statistics in samples from a normal population of two variables with matched sampling of one variable”. *Metron* 9:87–126.

Winship, Christopher/Stephen L. Morgan. 1999. “The estimation of causal effects from observational data”. *Annual Review of Sociology* 25 (1): 659–706.

Wong, Vivian C/Peter M Steiner. 2018. *Replication Designs for Causal Inference.* Tech. rep. 62.

Xie, Yu/Jennie E Brand/Ben Jann. 2012. “Estimating Heterogeneous Treatment Effects with Observational Data.” *Sociological methodology* 42 (1): 314–347.

Yoon, Jinsung/James Jordon/Mihaela van der Schaar. 2018. “GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets”. In *ICLR.*

Yu, Ling/Youpeng Zhong/Longsheng Bao/Zhanfei Wang. 2010. *The study on Shenyang Sanhao Bridge steel arch tala cable-stayed construction key technologies*, 26:292–295. 2.

Zhang, Jun/Kai F. Yu. 1998. “What’s the Relative Risk?” *JAMA* 280 (19): 1690.

Zhao, Zhong. 2004. “USING MATCHING TO ESTIMATE TREATMENT EFFECTS: DATA REQUIREMENTS, MATCHING METRICS, AND MONTE CARLO EVIDENCE”. *The Review of Economics and Statistics* 86 (1): 91–107.

Zigler, Corwin M./Francesca Dominici/Yun Wang. 2012. “Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes”. *Biostatistics* 13 (2): 289–302.

Zubizarreta, José R. 2012. “Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure After Surgery”. *Journal of the American Statistical Association* 107 (500): 1360–1371.

Zubizarreta, José R./Ricardo D. Paredes/Paul R. Rosenbaum. 2014. “Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile”. *Annals of Applied Statistics* 8 (1): 204–231. arXiv: `arXiv:1404.3584v1`.

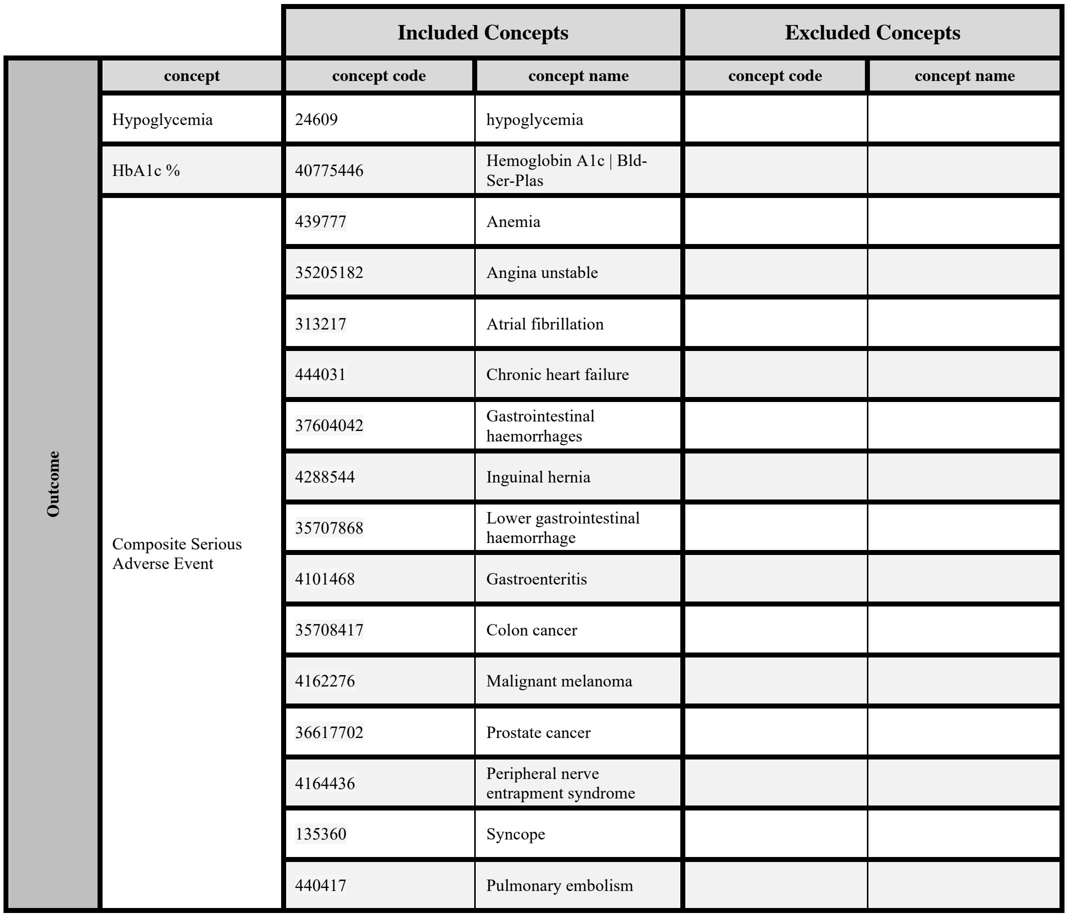
iSTAR Assessment. 2011. *Causal Reasoning.*

Chapter 8

Appendix

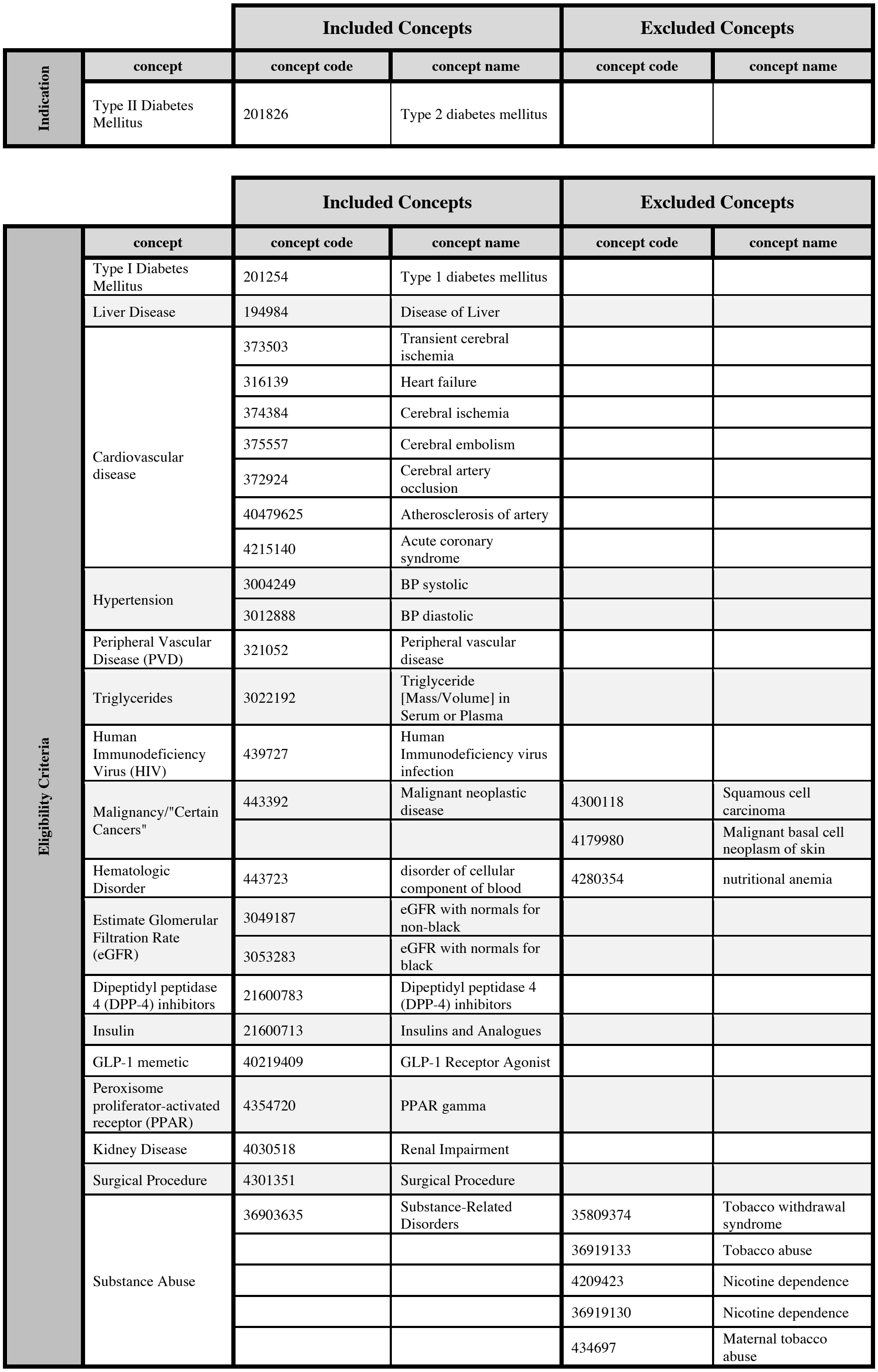
Appendix for Aim 1.1

Figure 8.1: Sitagliptin vs Glimepiride, Outcome Definitions



		Included Concepts		Excluded Concepts	
	concept	**concept code**	**concept name**	**concept code**	**concept name**
Outcome	Hypoglycemia	24609	hypoglycemia		
	HbA1c %	40775446	Hemoglobin A1c | Bld-Ser-Plas		
	Composite Serious Adverse Event	439777	Anemia		
		35205182	Angina unstable		
		313217	Atrial fibrillation		
		444031	Chronic heart failure		
		37604042	Gastrointestinal haemorrhages		
		4288544	Inguinal hernia		
		35707868	Lower gastrointestinal haemorrhage		
		4101468	Gastroenteritis		
		35708417	Colon cancer		
		4162276	Malignant melanoma		
		36617702	Prostate cancer		
		4164436	Peripheral nerve entrapment syndrome		
		135360	Syncope		
		440417	Pulmonary embolism		

Figure 8.2: Sitagliptin vs Glimepiride, Concept Definitions



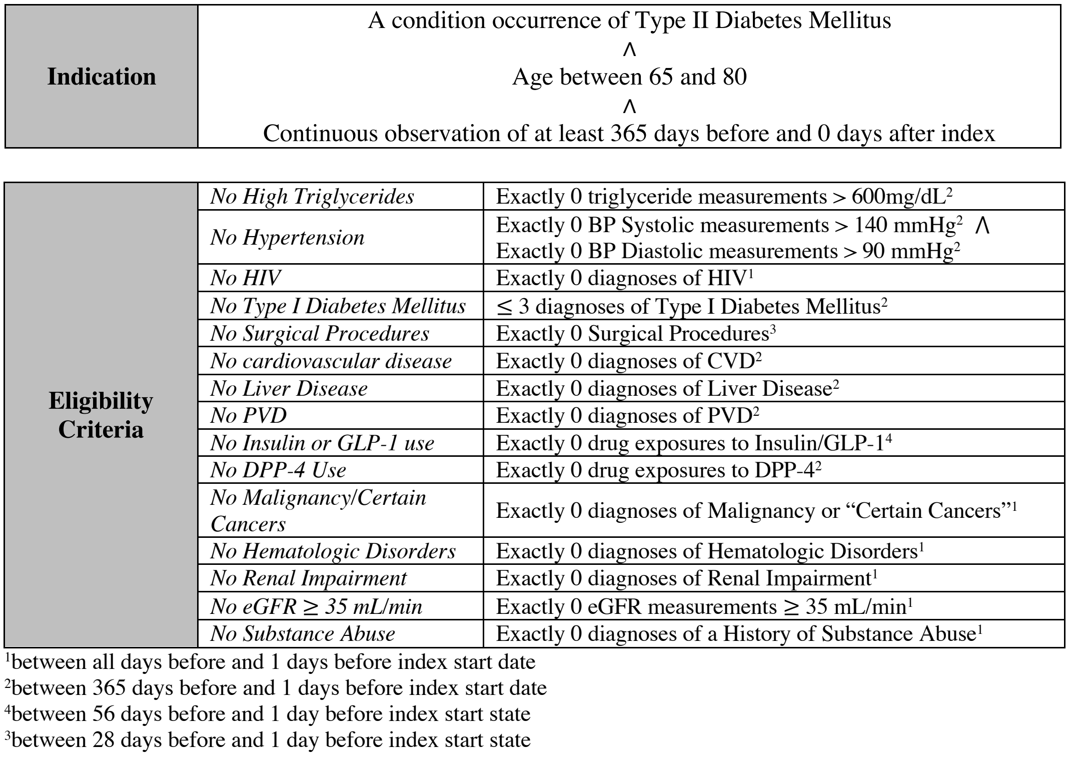
		Included Concepts		Excluded Concepts	
	concept	concept code	concept name	concept code	concept name
Indication	Type II Diabetes Mellitus	201826	Type 2 diabetes mellitus		

		Included Concepts		Excluded Concepts	
	concept	concept code	concept name	concept code	concept name
Eligibility Criteria	Type I Diabetes Mellitus	201254	Type 1 diabetes mellitus		
	Liver Disease	194984	Disease of Liver		
	Cardiovascular disease	373503	Transient cerebral ischemia		
		316139	Heart failure		
		374384	Cerebral ischemia		
		375557	Cerebral embolism		
		372924	Cerebral artery occlusion		
		40479625	Atherosclerosis of artery		
		4215140	Acute coronary syndrome		
	Hypertension	3004249	BP systolic		
		3012888	BP diastolic		
	Peripheral Vascular Disease (PVD)	321052	Peripheral vascular disease		
	Triglycerides	3022192	Triglyceride [Mass/Volume] in Serum or Plasma		
	Human Immunodeficiency Virus (HIV)	439727	Human Immunodeficiency virus infection		
	Malignancy/"Certain Cancers"	443392	Malignant neoplastic disease	4300118	Squamous cell carcinoma
				4179980	Malignant basal cell neoplasm of skin
	Hematologic Disorder	443723	disorder of cellular component of blood	4280354	nutritional anemia
	Estimate Glomerular Filtration Rate (eGFR)	3049187	eGFR with normals for non-black		
		3053283	eGFR with normals for black		
	Dipeptidyl peptidase 4 (DPP-4) inhibitors	21600783	Dipeptidyl peptidase 4 (DPP-4) inhibitors		
	Insulin	21600713	Insulins and Analogues		
	GLP-1 memetic	40219409	GLP-1 Receptor Agonist		
	Peroxisome proliferator-activated receptor (PPAR)	4354720	PPAR gamma		
	Kidney Disease	4030518	Renal Impairment		
	Surgical Procedure	4301351	Surgical Procedure		
	Substance Abuse	36903635	Substance-Related Disorders	35809374	Tobacco withdrawal syndrome
				36919133	Tobacco abuse
				4209423	Nicotine dependence
				36919130	Nicotine dependence
				434697	Maternal tobacco abuse

Note that all concepts definitions include descendants

Appendix for Aim 1.2

Figure 8.3: Sitagliptin vs Glimepiride, Cohort Creation



Indication	A condition occurrence of Type II Diabetes Mellitus ∧ Age between 65 and 80 ∧ Continuous observation of at least 365 days before and 0 days after index

Eligibility Criteria	*No High Triglycerides*	Exactly 0 triglyceride measurements > 600mg/dL[2]
	No Hypertension	Exactly 0 BP Systolic measurements > 140 mmHg[2] ∧ Exactly 0 BP Diastolic measurements > 90 mmHg[2]
	No HIV	Exactly 0 diagnoses of HIV[1]
	No Type I Diabetes Mellitus	≤ 3 diagnoses of Type I Diabetes Mellitus[2]
	No Surgical Procedures	Exactly 0 Surgical Procedures[3]
	No cardiovascular disease	Exactly 0 diagnoses of CVD[2]
	No Liver Disease	Exactly 0 diagnoses of Liver Disease[2]
	No PVD	Exactly 0 diagnoses of PVD[2]
	No Insulin or GLP-1 use	Exactly 0 drug exposures to Insulin/GLP-1[4]
	No DPP-4 Use	Exactly 0 drug exposures to DPP-4[2]
	No Malignancy/Certain Cancers	Exactly 0 diagnoses of Malignancy or "Certain Cancers"[1]
	No Hematologic Disorders	Exactly 0 diagnoses of Hematologic Disorders[1]
	No Renal Impairment	Exactly 0 diagnoses of Renal Impairment[1]
	No eGFR ≥ 35 mL/min	Exactly 0 eGFR measurements ≥ 35 mL/min[1]
	No Substance Abuse	Exactly 0 diagnoses of a History of Substance Abuse[1]

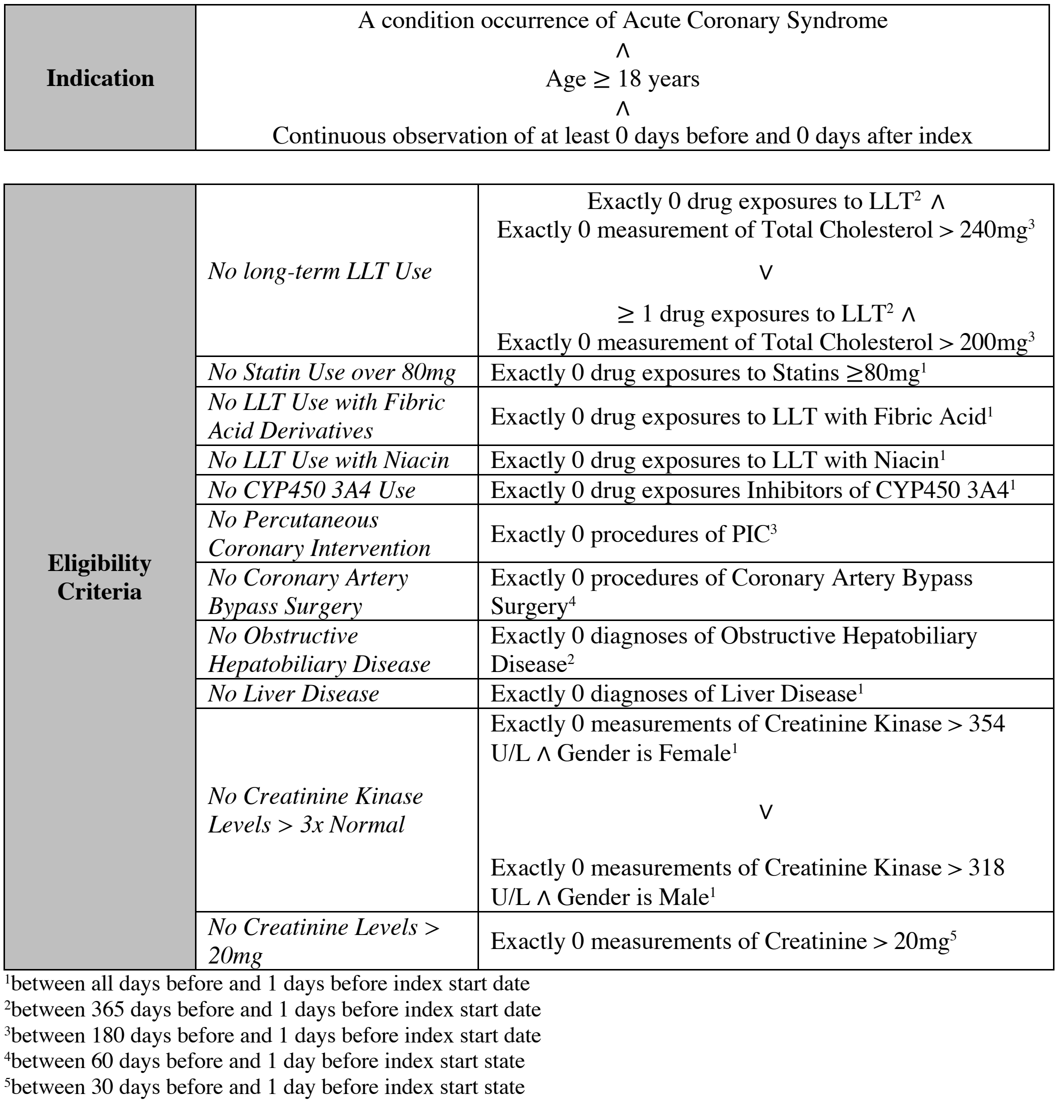
[1] between all days before and 1 days before index start date

[2] between 365 days before and 1 days before index start date

[4] between 56 days before and 1 day before index start state

[3] between 28 days before and 1 day before index start state

Figure 8.4: PROVE-IT, Cohort Creation



Indication	A condition occurrence of Acute Coronary Syndrome ∧ Age ≥ 18 years ∧ Continuous observation of at least 0 days before and 0 days after index

Eligibility Criteria	*No long-term LLT Use*	Exactly 0 drug exposures to LLT[2] ∧ Exactly 0 measurement of Total Cholesterol > 240mg[3] ∨ ≥ 1 drug exposures to LLT[2] ∧ Exactly 0 measurement of Total Cholesterol > 200mg[3]
	No Statin Use over 80mg	Exactly 0 drug exposures to Statins ≥80mg[1]
	No LLT Use with Fibric Acid Derivatives	Exactly 0 drug exposures to LLT with Fibric Acid[1]
	No LLT Use with Niacin	Exactly 0 drug exposures to LLT with Niacin[1]
	No CYP450 3A4 Use	Exactly 0 drug exposures Inhibitors of CYP450 3A4[1]
	No Percutaneous Coronary Intervention	Exactly 0 procedures of PIC[3]
	No Coronary Artery Bypass Surgery	Exactly 0 procedures of Coronary Artery Bypass Surgery[4]
	No Obstructive Hepatobiliary Disease	Exactly 0 diagnoses of Obstructive Hepatobiliary Disease[2]
	No Liver Disease	Exactly 0 diagnoses of Liver Disease[1]
	No Creatinine Kinase Levels > 3x Normal	Exactly 0 measurements of Creatinine Kinase > 354 U/L ∧ Gender is Female[1] ∨ Exactly 0 measurements of Creatinine Kinase > 318 U/L ∧ Gender is Male[1]
	No Creatinine Levels > 20mg	Exactly 0 measurements of Creatinine > 20mg[5]

[1] between all days before and 1 days before index start date

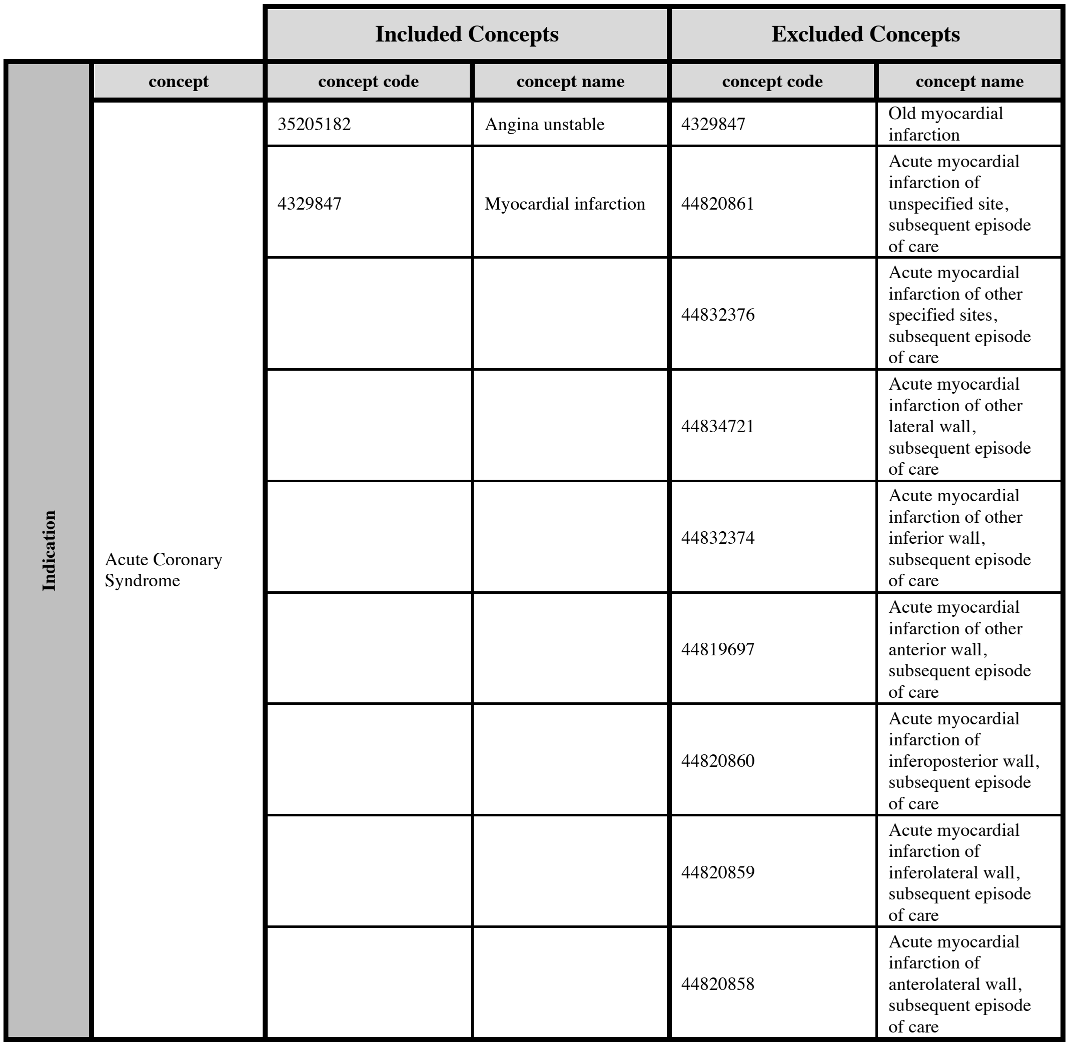
[2] between 365 days before and 1 days before index start date

[3] between 180 days before and 1 days before index start date

[4] between 60 days before and 1 day before index start state

[5] between 30 days before and 1 day before index start state

Figure 8.5: PROVE-IT, Concept Definitions



		Included Concepts		Excluded Concepts	
	concept	concept code	concept name	concept code	concept name
Indication	Acute Coronary Syndrome	35205182	Angina unstable	4329847	Old myocardial infarction
		4329847	Myocardial infarction	44820861	Acute myocardial infarction of unspecified site, subsequent episode of care
				44832376	Acute myocardial infarction of other specified sites, subsequent episode of care
				44834721	Acute myocardial infarction of other lateral wall, subsequent episode of care
				44832374	Acute myocardial infarction of other inferior wall, subsequent episode of care
				44819697	Acute myocardial infarction of other anterior wall, subsequent episode of care
				44820860	Acute myocardial infarction of inferoposterior wall, subsequent episode of care
				44820859	Acute myocardial infarction of inferolateral wall, subsequent episode of care
				44820858	Acute myocardial infarction of anterolateral wall, subsequent episode of care

		Included Concepts		Excluded Concepts	
	concept	**concept code**	**concept name**	**concept code**	**concept name**
Eligibility Criteria	Long Term Lipid-Lowering Therapy	21601853	Lipid Modifying Agents		
	Total Cholesterol	3027114	Cholesterol [Mass/volume] in Serum or Plasma		
	Statin	21601855	HMG CoA reductase inhibitors		
	Lipid Lowering Therapy with Fibric Acid	21601864	Fibrates		
	Lipid Lowering Therapy with Niacin	1517824	Niacin		
	CYP420-34	21601919	Imidazole and triazole derivatives		
	Percutaneous Coronary Intervention	4216130	Percutaneous coronary intervention		
	Coronary Artery Bypass Surgery	37522318	Coronary artery bypass		
	Obstructive Hepatobilliary Disease	35902850	Obstructive bile duct disorders (excl neoplasms)		
	Hepatic Disease	194984	Disease of liver		
	Creatinine Kinease Level	3007220	Creatine kinase [Enzymatic activity/volume] in Serum or Plasma		
	Creatinine Serum	3016723	Creatinine serum/plasma		

Note that all concepts definitions include descendants

Figure 8.6: ACCOMPLISH, Cohort Creation

Indication	A condition occurrence of Hypertension ∧ Age ≥ 55 years ∧ Continuous observation of at least 0 days before and 0 days after index

Eligibility Criteria	*High Systolic Blood Pressure or Treatment with Antihypertensives*	≥1 drug exposures to antihypertensive drug[1] ∨ ≥1 measurements of Systolic Blood Pressure ≥ 160mmHg[1]
	If Age ≥ 60 years, 1+ of the following; If 55 ≤ Age ≤ 60 years, 2+ of the following	
	Myocardial Infarction	a diagnosis of Myocardial Infarction[1]
	Unstable Angina	a diagnosis of Unstable Angina[1]
	Coronary Revascularization	A procedure of Coronary Revascularization[1]
	Stroke	A diagnosis of Stroke[1]
	Peripheral Arterial Occlusive Disease (PROC)	A diagnosis of PROC[1]
	Diabetes Mellitus	A diagnosis of Diabetes Mellitus[1]
	Left Ventricular Hypertrophy	A diagnosis of Left Ventricular Hypertrophy[1]
	Elevated Serum Creatinine	A measurement of Creatinine > 1.7[1] ∧ Gender is Male ∨ A measurement of Creatinine > 1.5[1] ∧ Gender is Female
	ACE Inhibitor or Aldosterone Receptor Blocker (ARB) Use with Elevated Albumin Creatinine Ratio	≥1 drug exposures to ACE Inhibitor/ARB drug[1] ∧ Albumin Creatinine Ratio > 300mg/dL[1] ∨ Exactly 0 drug exposures to ACE Inhibitor/ARB drug[1] ∧ Albumin Creatinine Ratio > 200mg/dL[1]
	No Angina Pectoris	Exactly 0 diagnoses of Angina Pectoris[2]
	No Heart Failure	Exactly 0 diagnoses of Myocardial Infarction[3]
	No Acute Coronary Syndrome	Exactly 0 diagnoses of ACS[3]
	No Coronary Revascularization	Exactly 0 procedures of coronary revascularization[2]
	No Stroke	Exactly 0 diagnoses of stroke[2]
	No Ischemic Cerebrovascular Episodes	Exactly 0 diagnoses of ischemic cerebrovascular episodes[2]

[1] between all days before and 1 days before index start date

[2] between 90 days before and 1 days before index start date

[2] between 30 days before and 1 days before index start date

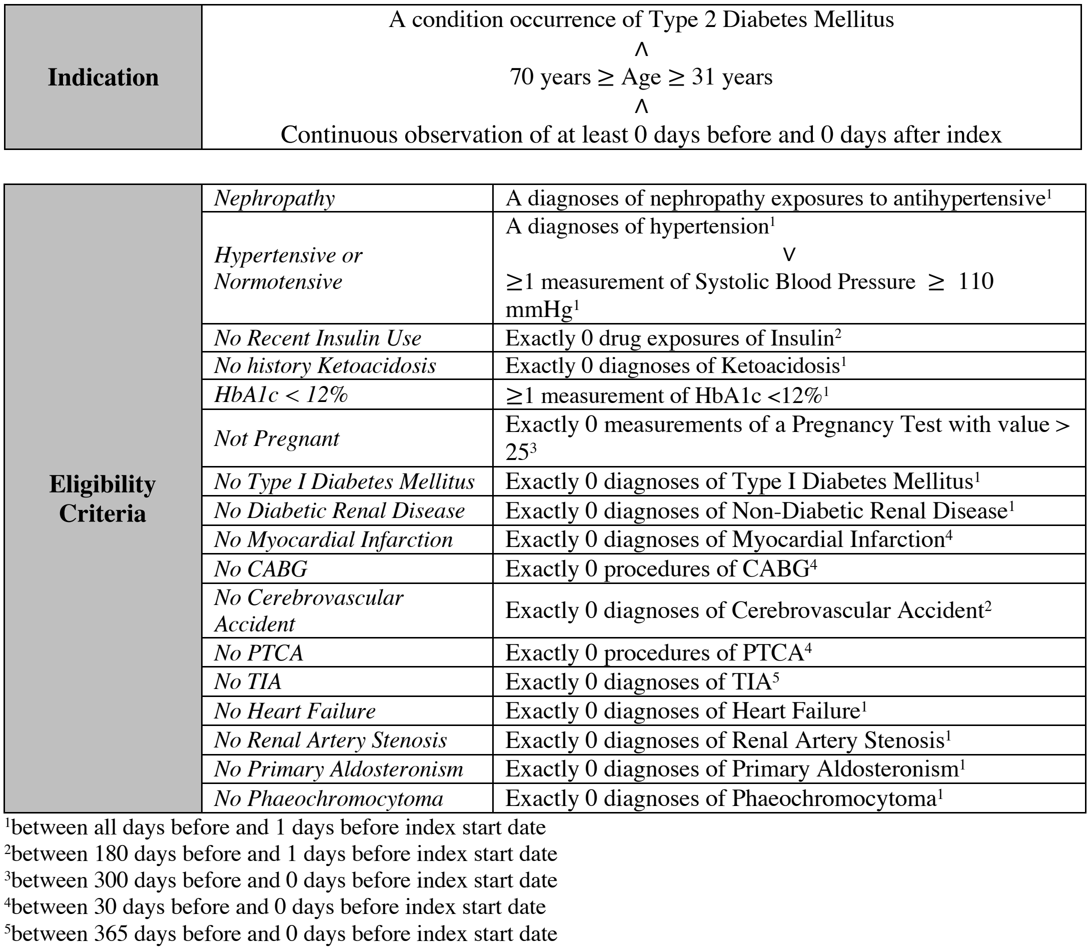
Figure 8.7: ACCOMPLISH, Concept Definitions

		Included Concepts		**Excluded Concepts**	
	concept	**concept code**	**concept name**	**concept code**	**concept name**
Indication	Hypertension	316866	Hypertensive disorder		

		Included Concepts		Excluded Concepts	
	concept	concept code	concept name	concept code	concept name
Eligibility Criteria	Antihypertensive Drug				
	ACE Inhibitors	21601783	Ace Inhibitors, Plain		
	Aldosterone	21601533	Aldosterone antagonists		
	Acute Coronary Syndrome	4215140	Acute coronary syndrome		
	Angina Pectoris	321318	Angina pectoris		
	Antihypertensive Drugs	21600381	Antihypertensives		
	Antidiabetic Drugs	21600744	Blood Glucose Lowering Drugs, Excluding Insulins		
		21600713	Insulins and Analogs		
		4336036	Oral Hypoglycemic Agents, Oral		
	Cardiovascular Disease (CVD)	134057	Disorder of cardiovascular system		
	Coronary Revascularization	37522318	Coronary artery bypass		
		4184298	Percutaneous transluminal angioplasty		
	Diabetes Mellitus	35502089	Glucose metabolism disorders (including diabetes mellitus)		
	Heart Failure	316139	Heart failure		
	Insulin	21600713	Insulins and Analogues		
	Ischemic Cerebrovascular Episodes	373503	Transient cerebral ischemia		
		36718067	Transient ischemic attack		
	Albumin	3024561	Albumin serum/plasma		
	Myocardial Infarctions	4329847	Myocardial infarction		
		35205189	Myocardial infarction		
	Overnight Fasting Plasma Glucose	3037110	Fasting glucose [Mass/volume] in Serum or Plasma		
	Peripheral Arterial Occlusive Disease (PROC)	2002187	Aorta-iliac-femoral bypass		
		37522314	Carotid endarterectomy		
		37522318	Coronary artery bypass		
		37520683	Leg amputation		
	Renal Disease	37019308	Renal disorder		
	Left Ventricular Hypertrophy	35205348	Ventricular hypertrophy	4231591	Right ventricular hypertrophy
	Serum Creatinine	3016723	Creatinine serum/plasma		
	Systolic blood pressure	3004249	BP systolic		
		3018586	Systolic blood pressure--sitting		
		3035856	Systolic blood pressure--standing		
		3009395	Systolic blood pressure--supine		
	Target Organ Damage	4349444	Hypertrophy, Left Ventricular		
		75650	Proteinuria		
		37019318	Renal failure		
		376103	Retinopathy		
		443605	Vascular dementia		
	Type 2 Diabetes Mellitus	201826	Type 2 diabetes mellitus		
	Unstable Angina	35205182	Angina unstable		

Note that all concepts definitions include descendants

Figure 8.8: RENAAL, Cohort Creation

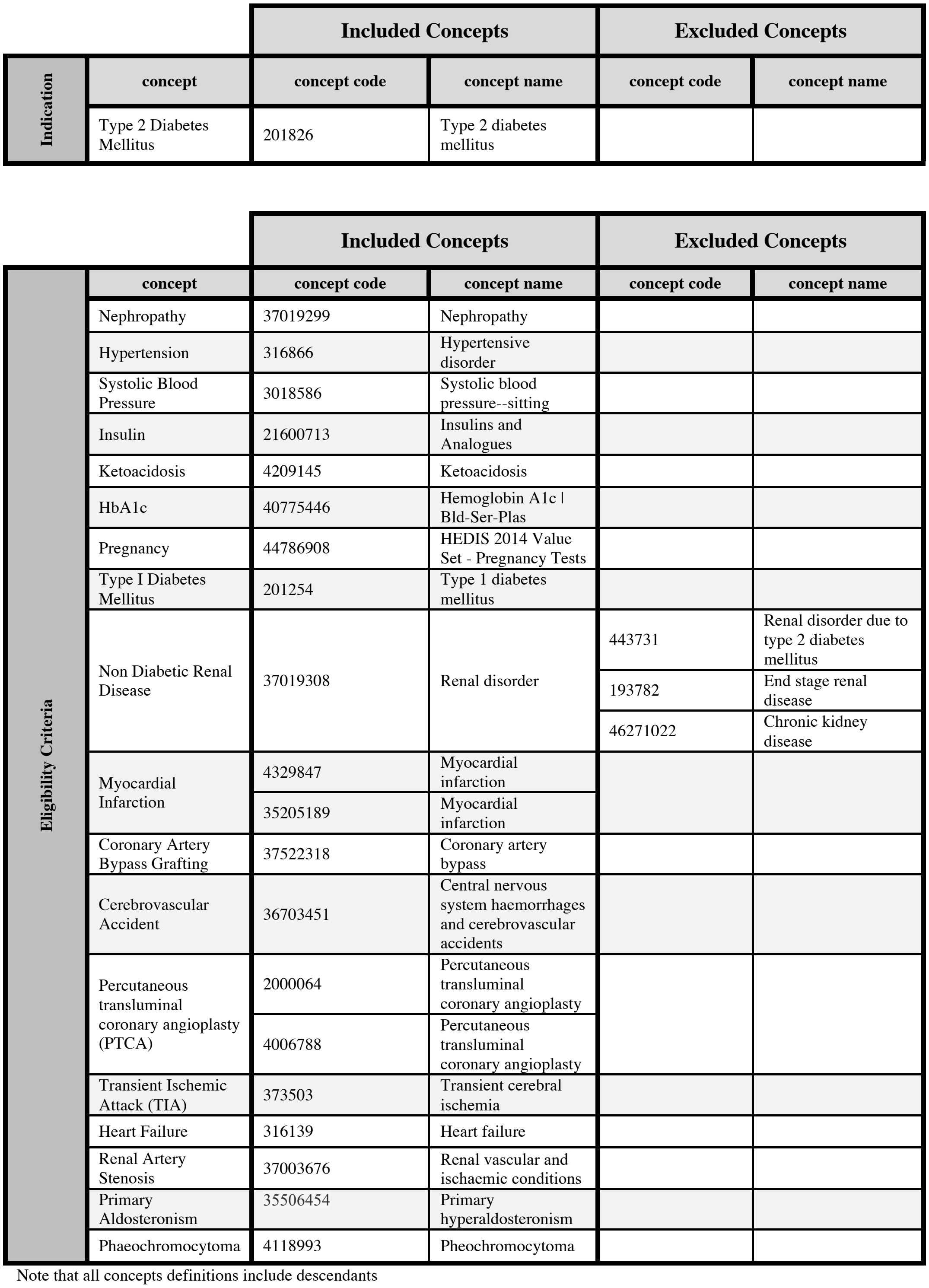


Indication	A condition occurrence of Type 2 Diabetes Mellitus $\wedge$ 70 years ≥ Age ≥ 31 years $\wedge$ Continuous observation of at least 0 days before and 0 days after index

Eligibility Criteria	*Nephropathy*	A diagnoses of nephropathy exposures to antihypertensive[1]
	Hypertensive or Normotensive	A diagnoses of hypertension[1] $\vee$ ≥1 measurement of Systolic Blood Pressure ≥ 110 mmHg[1]
	No Recent Insulin Use	Exactly 0 drug exposures of Insulin[2]
	No history Ketoacidosis	Exactly 0 diagnoses of Ketoacidosis[1]
	HbA1c < 12%	≥1 measurement of HbA1c <12%[1]
	Not Pregnant	Exactly 0 measurements of a Pregnancy Test with value > 25[3]
	No Type I Diabetes Mellitus	Exactly 0 diagnoses of Type I Diabetes Mellitus[1]
	No Diabetic Renal Disease	Exactly 0 diagnoses of Non-Diabetic Renal Disease[1]
	No Myocardial Infarction	Exactly 0 diagnoses of Myocardial Infarction[4]
	No CABG	Exactly 0 procedures of CABG[4]
	No Cerebrovascular Accident	Exactly 0 diagnoses of Cerebrovascular Accident[2]
	No PTCA	Exactly 0 procedures of PTCA[4]
	No TIA	Exactly 0 diagnoses of TIA[5]
	No Heart Failure	Exactly 0 diagnoses of Heart Failure[1]
	No Renal Artery Stenosis	Exactly 0 diagnoses of Renal Artery Stenosis[1]
	No Primary Aldosteronism	Exactly 0 diagnoses of Primary Aldosteronism[1]
	No Phaeochromocytoma	Exactly 0 diagnoses of Phaeochromocytoma[1]

[1] between all days before and 1 days before index start date
[2] between 180 days before and 1 days before index start date
[3] between 300 days before and 0 days before index start date
[4] between 30 days before and 0 days before index start date
[5] between 365 days before and 0 days before index start date

Figure 8.9: RENAAL, Concept Definitions



		Included Concepts		Excluded Concepts	
	concept	concept code	concept name	concept code	concept name
Indication	Type 2 Diabetes Mellitus	201826	Type 2 diabetes mellitus		

		Included Concepts		Excluded Concepts	
	concept	concept code	concept name	concept code	concept name
Eligibility Criteria	Nephropathy	37019299	Nephropathy		
	Hypertension	316866	Hypertensive disorder		
	Systolic Blood Pressure	3018586	Systolic blood pressure--sitting		
	Insulin	21600713	Insulins and Analogues		
	Ketoacidosis	4209145	Ketoacidosis		
	HbA1c	40775446	Hemoglobin A1c | Bld-Ser-Plas		
	Pregnancy	44786908	HEDIS 2014 Value Set - Pregnancy Tests		
	Type I Diabetes Mellitus	201254	Type 1 diabetes mellitus		
	Non Diabetic Renal Disease	37019308	Renal disorder	443731	Renal disorder due to type 2 diabetes mellitus
				193782	End stage renal disease
				46271022	Chronic kidney disease
	Myocardial Infarction	4329847	Myocardial infarction		
		35205189	Myocardial infarction		
	Coronary Artery Bypass Grafting	37522318	Coronary artery bypass		
	Cerebrovascular Accident	36703451	Central nervous system haemorrhages and cerebrovascular accidents		
	Percutaneous transluminal coronary angioplasty (PTCA)	2000064	Percutaneous transluminal coronary angioplasty		
		4006788	Percutaneous transluminal coronary angioplasty		
	Transient Ischemic Attack (TIA)	373503	Transient cerebral ischemia		
	Heart Failure	316139	Heart failure		
	Renal Artery Stenosis	37003676	Renal vascular and ischaemic conditions		
	Primary Aldosteronism	35506454	Primary hyperaldosteronism		
	Phaeochromocytoma	4118993	Pheochromocytoma		

Note that all concepts definitions include descendants

Appendix for Aim 2.1

Results of the sensitivity analysis for the Counterfactual χ-GAN (cGAN) simulation on average treatment effect (ATE). Investigating cGAN and comparator performance on ATE estimation as a function of (i) the per-arm sample size (N); (ii) the unbiased average treatment effect that exists in the truly counterfactual populations (ATE); and (iii) the size of the truly counterfactual populations as a proportion of the total population (overlap). Blanks denote parameter combinations where the method failed.

	overlap = 0.1								
	N=2000			N=4000			N=8000		
	ATE=400	ATE=70	ATE=0.2	ATE=400	ATE=70	ATE=0.2	ATE=400	ATE=70	ATE=0.2
unweighted	-320.02	33.94	2.00	-320.01	34.02	2.02	-319.99	34.01	1.99
cGAN	**380.84**	**68.59**	**0.27**	**397.11**	**69.99**	**0.40**	**393.93**	**69.84**	**0.20**
IPW	-243.06	48.65	2.14	-263.69	44.08	2.06	-262.37	44.67	2.03
clipped IPW	-315.23	35.86	2.03	-317.81	34.66	2.03	-318.35	34.67	1.99
PS ATE	-243.06	48.65	2.14	-263.69	44.08	2.06	-262.37	44.67	2.03
Twang ATE	-291.83	41.05	2.08	-287.34	40.86	2.08	-283.29	41.21	2.03
CBPS ATE	-244.10	48.47	2.14	-268.38	43.47	2.07	-264.68	44.39	2.03
NPCBPS ATE	-162.12	10.62	0.76	-28.95	12.98	-0.09	70.66	38.77	0.93
Ebal ATE	-	-	-	-	-	-	-	-	-
EBCW ATE	-	-	-	-	-	-	-	-	-
OptWeight ATE	-121.22	65.84	3.09	48.33	95.04	3.48	-481.75	59.62	-0.29

	overlap = 0.5								
	N=2000			N=4000			N=8000		
	ATE=400	ATE=70	ATE=0.2	ATE=400	ATE=70	ATE=0.2	ATE=400	ATE=70	ATE=0.2
unweighted	-0.05	50.02	1.15	-0.02	49.98	1.23	0.00	50.00	1.21
cGAN	**395.98**	**69.84**	**0.15**	**398.84**	**69.85**	**0.22**	**396.34**	**69.84**	**0.23**
IPW	303.49	109.02	1.34	245.04	95.42	1.38	234.96	94.21	1.45
clipped IPW	204.59	91.74	1.30	201.40	87.92	1.37	202.73	88.48	1.42
PS ATE	303.49	109.02	1.34	245.04	95.42	1.38	234.96	94.21	1.45
Twang ATE	186.64	84.48	1.25	192.12	84.87	1.36	195.88	84.86	1.33
CBPS ATE	255.46	101.84	1.32	230.91	93.46	1.38	227.68	93.51	1.47
NPCBPS ATE	30.56	26.88	0.43	-41.37	17.46	0.02	226.39	48.89	1.96
Ebal ATE	44.73	101.65	2.18	151.19	104.79	1.91	272.24	114.02	1.76
EBCW ATE	22.36	50.83	1.09	75.59	52.40	0.96	136.24	57.01	0.88
OptWeight ATE	17.97	50.73	1.10	77.52	52.37	0.96	129.05	56.93	0.82

	overlap = 0.9								
	N=2000			N=4000			N=8000		
	ATE=400	ATE=70	ATE=0.2	ATE=400	ATE=70	ATE=0.2	ATE=400	ATE=70	ATE=0.2
unweighted	320.03	65.99	0.39	319.98	65.97	0.42	320.00	66.01	0.39
cGAN	**391.30**	**69.70**	**0.21**	**398.60**	**69.87**	**0.22**	**399.23**	**69.97**	**0.19**
IPW	687.17	133.96	0.69	687.76	132.10	0.64	688.01	133.43	0.65
clipped IPW	667.27	132.96	0.71	663.66	130.67	0.69	669.53	132.36	0.69
PS ATE	667.17	133.96	0.69	687.76	132.10	0.64	688.01	133.43	0.65
Twang ATE	660.29	126.71	0.59	663.88	126.30	0.61	664.90	126.73	0.56
CBPS ATE	677.92	133.21	0.72	678.58	131.59	0.68	675.98	132.44	0.67
NPCBPS ATE	126.94	30.07	0.22	161.38	32.48	0.18	159.74	30.11	0.12
Ebal ATE	608.81	131.94	0.91	663.72	132.23	0.75	649.80	132.49	0.78
EBCW ATE	304.40	65.97	0.45	331.87	66.12	0.38	324.90	66.24	0.39
OptWeight ATE	303.84	65.96	0.45	331.56	66.08	0.37	324.93	66.25	0.39

Results of the sensitivity analysis for the Counterfactual χ-GAN (cGAN) simulation on Estimated Sample Size (ESS). Investigating cGAN and comparator performance on ESS estimation as a function of (i) the per-arm sample size (N) and (ii) the size of the truly counterfactual populations as a proportion of the total population (overlap). A high-quality ESS is one which converges with $(N * 2) * overlap$. This should approximate the number of units over which counterfactual inference is appropriate. Note that ESS does not change over variations in true average treatment effect. Blanks denote parameter combinations where the method failed.

	overlap = 0.1			overlap = 0.5			overlap = 0.9		
	N=2000	N=4000	N=8000	N=2000	N=4000	N=8000	N=2000	N=4000	N=8000
unweighted	4000.00	8000.00	16000.00	4000.00	8000.00	16000.00	4000.00	8000.00	16000.00
cGAN	**389.69**	**755.53**	**1520.70**	**1928.02**	**3913.03**	**7896.35**	**3548.88**	**7126.55**	**14276.81**
IPW	1632.30	6106.11	12101.71	2153.86	5832.96	12741.01	3807.50	7650.75	15341.37
clipped IPW	3980.70	7996.73	15981.06	3351.64	6838.98	13881.66	3884.54	7805.16	15611.48
PS ATE	1632.30	6106.11	12101.71	2153.86	5831.96	12741.01	3807.50	7650.73	15341.37
Twang ATE	3550.66	7436.22	14876.05	3590.72	7064.62	14425.47	3866.83	7745.95	15539.17
CBPS ATE	1658.62	6489.88	12490.85	2608.01	6167.73	13140.15	3878.17	7789.09	15587.93
NPCBPS ATE	12.83	5.00	5.04	85.10	36.42	2.22	74.32	6943.35	47.38
Ebal ATE	-	-	-	458.70	323.47	164.66	2590.64	6336.13	13867.72
EBCW ATE	-	-	-	458.70	323.46	164.66	2590.64	6336.07	13867.72
OptWeight ATE	10.68	10.58	9.58	828.62	594.60	275.50	2796.29	6591.70	14041.60

Appendix for Aim 3.1

The full joint distribution for the Noisy-OR Risk Allocation (NORA) model is given by Equation 8.1

$$p(Z, Y, R, \alpha, \beta) = p(\alpha; \lambda)p(\beta; \kappa) \prod^{N} p(Y_n|Z_{n,1:K}) \prod^{K} p(R_k|\alpha, \beta) \prod^{K} \prod^{N} p(Z_{n,k}|X_{n,k}, R_k) \tag{8.1}$$

Posterior of Z The posterior of Z is given by,

$$p(Z|X, Y, R) = \frac{p(R, X, Y|Z)p(Z)}{p(R, X, Y)}$$

The likelihood of Z is $p(R, X, Y|Z)$ and can be determined upon the removal of terms from the full joint that are independent of Z.

$$p(X, Y, R|Z) = \prod^{N} p(Y_n|Z_{n,1:K})$$

The prior, $p(Z)$, and the likelihood, $p(X, Y, R|Z)$ are given by the functional forms,

$$p(Z_{n,k}|X_{n,k}R_k) = X_{n,k}R_k^{Z_{n,k}}(1 - X_{n,k}R_k)^{1-Z_{n,k}}$$

$$p(Y_n|Z_{1:K}) = \big[1 - \prod_{k=1}^{K} \big(1 - Z_{n,k}\big)\big]^{Y_n} \big[\prod_{k=1}^{K} \big(1 - Z_{n,k}\big)\big]^{(1-Y_n)}$$

Therefore, the posterior of Z is given by,

$$p(Z|R, X, Y) = \prod^{N} p(Y_n|Z_{n,1:K}) \prod^{N} \prod^{K} \Big[p(Z_{n,k}|X_{n,k}R_k)\Big]$$

We can recover the probability distribution of Z from the joint distribution by summing over the discrete cases of Z and removing the those ns which do not contribute.

$$p(Z_{n,k}|\ldots) \propto \big[1 - \big(1 - Z_{n,k}\big) \prod_{j \neq k} \big(1 - Z_{n,j}\big)\big]^{Y_n} \big[\big(1 - Z_{n,k}\big) \prod_{j \neq k} \big(1 - Z_{n,j}\big)\big]^{(1-Y_n)} \\ (X_{n,k}R_k)^{Z_{n,k}} (1 - X_{n,k}R_k)^{1-Z_{n,k}} \tag{8.2}$$

Posterior of R The posterior of R is given by,

$$p(R|Z, X, Y) = \frac{p(Z, X, Y|R)p(R)}{p(Z, X, Y)} \tag{8.3}$$

The likelihood of R, $p(Z, X, Y|R)$, can be determined upon the removal of terms from the full joint that are independent of R.

$$p(R|Z, X, Y) = \prod^{K} p(R_k|\alpha, \beta) \prod^{N} \prod^{K} \Big[p(z_{n,k}|x_{n,k}R_k)\Big] \tag{8.4}$$

Plug in functional form of the expressions above, which are given by the following,

$$p(R_k|\alpha,\beta) = Beta(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\Big[R_k^{(\alpha-1)}(1-R_k)^{(\beta-1)}\Big] \tag{8.5}$$

$$p(z_{n,k}|X_{n,k},R_k) = (X_{n,k}R_k)^{Z_{n,k}}(1-X_{n,k}R_k)^{(1-Z_{n,k})} \tag{8.6}$$

Therefore, the posterior of R is given by,

$$p(R|Z,X,Y) = \frac{1}{B(\alpha,\beta)}\Bigg[\prod^{K}\Big[R_k^{(\alpha-1)}(1-R_k)^{(\beta-1)}\Big]\prod^{N}\prod^{K}\Big[(X_{n,k}R_k)^{Z_{n,k}}(1-X_{n,k}R_k)^{(1-Z_{n,k})}\Big]\Bigg] \tag{8.7}$$

Breaking exposures into k and $\neg k$,

$$p(R|Z,X,Y) = \frac{1}{B(\alpha,\beta)}\Big[R_k^{(\alpha-1)}(1-R_k)^{(\beta-1)}\Big]\Big[R_{\neg_k}^{(\alpha-1)}(1-R_{\neg_k})^{(\beta-1)}\Big]$$
$$\prod^{N}\Big[(X_{n,k}R_k)^{Z_{n,k}}(1-X_{n,k}R_k)^{(1-Z_{n,k})}\Big]\Big[(X_{n,\neg k}R_{\neg k})^{Z_{n,\neg k}}(1-X_{n,\neg k}R_{\neg k})^{(1-Z_{n,\neg k})}\Big] \tag{8.8}$$

Pull out exposures, $\neg k$, that are unrelated to, k.

$$p(R|Z,X,Y) = \frac{1}{B(\alpha,\beta)}\Big[R_k^{(\alpha-1)}(1-R_k)^{(\beta-1)}\Big]\prod^{N}\Big[(X_{n,k}R_k)^{Z_{n,k}}(1-X_{n,k}R_k)^{(1-Z_{n,k})}\Big] \tag{8.9}$$

Distribute exponent out over X and R, and then distribute out the product

$$p(R|Z,X,Y) = \frac{1}{B(\alpha,\beta)}\Big[R_k^{(\alpha-1)}(1-R_k)^{(\beta-1)}\Big]\prod^{N}\Big[X_{n,k}^{Z_{n,k}}\Big]\prod^{N}\Big[R_k^{Z_{n,k}}\Big]\prod^{N}\Big[(1-X_{n,k}R_k)^{(1-Z_{n,k})}\Big] \tag{8.10}$$

Rearrange to group like terms; product in exponent will become sum by virtue of a shared base and Z taking a value of either 0 or 1.

$$p(R|Z,X,Y) = \frac{1}{B(\alpha,\beta)}\left[R_k^{(\alpha-1)}(1-R_k)^{(\beta-1)}\right]\left[R_k^{\sum^N Z_{n,k}}\right]\prod^N\left[X_{n,k}^{Z_{n,k}}\right]\prod^N\left[(1-X_{n,k}R_k)^{(1-Z_{n,k})}\right] \tag{8.11}$$

Because $X_{n,k}$ is either 0 or 1, and Z is either 0 or 1, $X_{n,k}^Z$ will always evaluate to 1, as $1^1 = 1$, $1^0 = 1$, $0^1 = 1$, and $0^0 = 1$. Therefore the expression $X_{n,k}^Z$ drops out of the posterior.

$$p(R|Z,X,Y) = \frac{1}{B(\alpha,\beta)}\left[R_k^{(\alpha-1)}(1-R_k)^{(\beta-1)}\right]\left[R_k^{\sum^N Z_{n,k}}\right]\prod^N\left[(1-X_{n,k}R_k)^{(1-Z_{n,k})}\right] \tag{8.12}$$

Similarly, in the expression $(1-X_{n,k}R_k)^{(1-Z)}$, when X is 0, the expression simplifies to $1^{(1-Z)}$; which will either be $1^1 = 1$ or $1^0 = 1$ under the different values of Z. Therefore, in this case where X is 0, the expression drops out of the posterior. When X is 1, the expression simplifies to $(1-R_k)^{(1-Z)}$. This simplified expression with make a contribution of the total number of times the exposure k occurs across all patients

$$p(R|Z,X,Y) = \frac{1}{B(\alpha,\beta)}\left[R_k^{(\alpha-1)}(1-R_k)^{(\beta-1)}\right]\left[R_k^{\sum^N Z_{n,k}}\right]\left[(1-R_k)^{(1-\sum^N Z_{n,k})}\right] \tag{8.13}$$

Group exponents according to shared bases.

$$p(R|Z,X,Y) = \frac{1}{B(\alpha,\beta)}\left[(R_k^{(\alpha-1+\sum^N Z_k)})(1-R_k)^{(\beta-1-\sum^N X_k-\sum^N Z_k)}\right] \tag{8.14}$$

which is equivalent to a Beta distribution, $Beta(\alpha^*, \beta^*)$, where α^* and β^* are;

$$\alpha^* = \alpha + \sum^{N} Z_k \tag{8.15}$$

$$\beta^* = \beta + \sum^{N} X_k - \sum^{N} Z_k \tag{8.16}$$

Appendix for Aim 3.2

Table 8.1: DIC Procedure

	NORA		**L1**		**Levin AR**		**RR**		**GPS-EBGM**	
Rank	Concept	Risk	Concept	Risk (OR)	Concept	Risk	Concept	Risk (RR)	Concept	Risk (EBGM)
1	Continuous invasive mechanical ventilation for 96 consecutive hours or more	0.0310	Transfusion of packed cells	0.6866 (2.19)	Transfusion of packed cells	0.3913	Bronchoscopy	0.9986 (701.55)	Transfusion of other serum	0.9760 (41.66)
2	Transfusion of platelets	0.0235	Insertion of endotracheal tube	0.6403 (1.78)	Insertion of endotracheal tube	0.3177	Open chest cardiac massage	0.9986 (701.55)	Other repair of chest wall	0.9733 (37.49)
3	Hemodialysis	0.0109	intercept	0.0015 (0.00)	Injection of antibiotic	0.2568	Renal autotransplantation	0.9986 (701.55)	Transfusion of platelets	0.9699 (33.22)
4	Mastectomy, partial	0.0076			Venous catheterization, not elsewhere classified	0.2522	Resection of vessel with anastomosis, abdominal arteries	0.9986 (701.55)	Continuous invasive mechanical ventilation for 96 consecutive hours or more	0.9659 (29.31)
5	Bronchoscopy	0.0062			Continuous invasive mechanical ventilation for 96 consecutive hours or more	0.2520	Interatrial transposition of venous return	0.9972 (350.77)	Arterial catheterization	0.9651 (28.64)

Table 8.2: Glaucoma

	NORA		L1		Levin AR		RR		GPS-EBGM	
Rank	Concept	Risk	Concept	Risk (OR)	Concept	Risk	Concept	Risk (RR)	Concept	Risk (EBGM)
1	Blind hypertensive eye	0.4823	Secondary hypertension	0.6930 (2.26)	Essential hypertension	0.2843	Acute osteomyelitis of humerus	0.9570 (23.28)	Blind hypertensive eye	0.8955 (9.57)
2	Complication	0.2068	Type 2 diabetes mellitus	0.6386 (1.77)	Type 2 diabetes mellitus	0.1754	Acute parainfluenza virus bronchitis	0.9570 (23.28)	Complication	0.8185 (5.51)
3	Secondary hypertension	0.1291	Essential hypertension	0.6215 (1.64)	Osteoarthritis	0.1267	Adverse reaction to central nervous system muscle-tone depressants	0.9570 (23.28)	Secondary hypertension	0.7955 (4.89)
4	Mature cataract	0.1022	Osteoarthritis	0.5869 (1.42)	Type II diabetes mellitus uncontrolled	0.1131	Adverse reaction to mixed bacterial vaccine	0.9570 (23.28)	Retinal detachment	0.7863 (4.68)
5	Retinal detachment	0.0948	Age-related cataract	0.5743 (1.35)	Pure hypercholesterolemia	0.1102	Adverse reaction to substance	0.9570 (23.28)	Buphthalmos	0.7812 (4.57)

Table 8.3: Hearing Loss

	NORA		L1		Levin AR		RR		GPS-EBGM	
Rank	Concept	Risk	Concept	Risk (OR)	Concept	Risk	Concept	Risk (RR)	Concept	Risk (EBGM)
1	carbamide peroxide	0.0259	carbamide peroxide	0.7148 (2.51)	Omeprazole	0.0332	denileukin diftitox	0.9889 (90.07)	carbamide peroxide	0.7033 (3.37)
2	intercept	0.0157	Alendronate	0.6199 (1.63)	Hydrochlorothiazide	0.0330	Diethylstilbestrol	0.9889 (90.07)	Flurbiprofen	0.6764 (3.09)
3	Alendronate	0.0111	Omeprazole	0.6195 (1.63)	Alendronate	0.0239	fidaxomicin	0.9889 (90.07)	Alendronate	0.5902 (2.44)
4	pneumococcal capsular polysaccharide type 7F vaccine	$4.87e^{-17}$	Flurbiprofen	0.6148 (1.60)	Metformin	0.0229	Dipivefrin	0.9667 (30.02)	Isoetharine	0.5885 (2.43)
5	Neomycin	$2.83e^{-18}$	topiramate	0.6132 (1.59)	carbamide peroxide	0.0199	Interferon Alfa-2a	0.9667 (30.02)	Urea	0.5816 (2.39)

Table 8.4: Heart Failure

	NORA		L1		Levin AR		RR		GPS-EBGM	
Rank	Concept	Risk	Concept	Risk (OR)	Concept	Risk	Concept	Risk (RR)	Concept	Risk (EBGM)
1	Cardiomyopathy	0.3587	Atrial fibrillation	0.9461 (2.87)	Essential hypertension	0.4399	Accidental poisoning by liquefied petroleum gas distributed in mobile containers	0.9346 (15.28)	Diseases of mitral and aortic valves	0.8711 (7.76)
2	Mitral valve stenosis	0.2821	Cardiomyopathy	0.9389 (2.73)	Type 2 diabetes mellitus	0.2111	Acquired forearm deformity	0.9346 (15.28)	Cardiomyopathy	0.8670 (7.52)
3	Rheumatic mitral regurgitation	0.2644	Essential hypertension	0.9023 (2.22)	Chest pain	0.1622	Acute peptic ulcer without hemorrhage AND without perforation	0.9346 (15.28)	Rheumatic mitral regurgitation	0.8667 (7.5)
4	Acute pulmonary edema	0.2626	Secondary hypertension	0.8940 (2.13)	Pure hypercholesterolemia	0.1568	Adhesions of iris	0.9346 (15.28)	Peripheral circulatory disorder associated with type 1 diabetes mellitus	0.8494 (6.64)
5	Acute myocardial infarction of anterior wall	0.2579	Chronic ischemic heart disease	0.8880 (2.07)	Coronary arteriosclerosis in native artery	0.1379	Adverse reaction to central nervous system muscle-tone depressants	0.9346 (15.28)	Mitral valve stenosis	0.8487 (6.61)

Table 8.5: Kaposi Sarcoma

	NORA		L1		Levin AR		RR		GPS-EBGM	
Rank	Concept	Risk	Concept	Risk (OR)	Concept	Risk	Concept	Risk (RR)	Concept	Risk (EBGM)
1	Human immunodeficiency virus infection	0.0070	Primary malignant neoplasm of fallopian tube	0.8428 (5.36)	Essential hypertension	0.2196	Carcinoma in situ of liver and/or biliary system	0.9979 (482.0)	Human immunodeficiency virus infection	0.9603 (25.2)
2	Malignant neoplasm of skin	0.0056	Secondary malignant neoplasm of female genital organ	0.8017 (4.04)	Human immunodeficiency virus infection	0.2022	Hodgkin's disease of lymph nodes of axilla AND/OR upper limb	0.9979 (482.0)	Malignant neoplasm of corpus uteri, excluding isthmus	0.8985 (9.85)
3	Primary malignant neoplasm of upper limb	0.0048	Human immunodeficiency virus infection	0.7872 (3.70)	Primary malignant neoplasm of female breast	0.1353	Primary malignant neoplasm of upper limb	0.9979 (482.0)	Pulmonary tuberculosis	0.5305 (2.13)
4	Malignant neoplasm of corpus uteri, excluding isthmus	0.0042	Primary malignant neoplasm of ovary	0.7778 (3.50)	Primary malignant neoplasm of unspecified site	0.1173	Transsexual	0.9979 (482.0)	Postprocedural pelvic peritoneal adhesions	0.1453 (1.17)
5	Secondary malignant neoplasm of female genital organ	0.0040	Hodgkin's disease of lymph nodes of axilla AND/OR upper limb	0.5649 (1.30)	Asymptomatic human immunodeficiency virus infection	0.0997	Subacute delirium	0.9974 (385.6)	Asymptomatic human immunodeficiency virus infection	0.1071 (1.12)

Table 8.6: Mucositis vs Ingredient-Level Drugs

	NORA		**L1**		**Levin AR**		**RR**		**GPS-EBGM**	
Rank	Concept	Risk	Concept	Risk (OR)	Concept	Risk	Concept	Risk (RR)	Concept	Risk (EBGM)
1	intercept	0.0908	Loratadine	0.7253 (2.64)	Loratadine	0.0454	clofarabine	0.9401 (16.69)	Mometasone	0.6667 (3.0)
2	Loratadine	0.0716	fluticasone	0.6168 (1.61)	fluticasone	0.0237	difenoxin	0.9401 (16.69)	olopatadine	0.6587 (2.93)
3	Mometasone	0.0620	Pseudoephedrine	0.5138 (1.06)	Pseudoephedrine	0.0160	trovafloxacin	0.9401 (16.69)	Loratadine	0.6441 (2.81)
4	Pseudoephedrine	0.0228	pneumococcal capsular polysaccharide vaccines	0.4964 (0.99)	Mometasone	0.0158	Astemizole	0.9201 (12.52)	irbesartan	0.6296 (2.7)
5	Cisapride	$3.69e^{-22}$	Influenza A virus vaccine; Influenza B virus vaccine, B-Massachusetts-2-2012-like virus	0.4937 (0.99)	montelukast	0.0103	Nedocromil	0.9201 (12.52)	Pseudoephedrine	0.6047 (2.53)

Table 8.7: Renal Impairment

	NORA		L1		Levin AR		RR		GPS-EBGM	
Rank	Concept	Risk	Concept	Risk (OR)	Concept	Risk	Concept	Risk (RR)	Concept	Risk (EBGM)
1	Disorder of transplanted kidney	0.4774	Essential hypertension	0.7561 (3.10)	Essential hypertension	0.5301	Abdominal rigidity of left lower quadrant	0.9299 (14.27)	Diabetic renal disease	0.8828 (8.53)
2	Chronic glomerulonephritis	0.3789	Disorder of kidney and/or ureter	0.6980 (2.31)	Type 2 diabetes mellitus	0.2697	Abnormal jaw closure	0.9299 (14.27)	Chronic glomerulonephritis	0.8630 (7.3)
3	Nephrotic syndrome	0.3775	Congestive heart failure	0.6979 (2.31)	Pure hypercholesterolemia	0.2176	Accidental poisoning by carbon monoxide...	0.9299 (14.27)	Disorder of transplanted kidney	0.8561 (6.95)
4	Hypertensive renal disease	0.3701	Type 2 diabetes mellitus	0.6484 (1.84)	Congestive heart failure	0.1547	Accidental poisoning by liquefied petroleum gas distributed in mobile containers	0.9299 (14.27)	Nephrotic syndrome	0.8559 (6.94)
5	Diabetic renal disease	0.3258	Secondary hypertension	0.5908 (1.44)	Type II diabetes mellitus uncontrolled	0.1462	Acquired claw foot	0.9299 (14.27)	Peripheral circulatory disorder associated with type 1 diabetes mellitus	0.8538 (6.84)

Table 8.8: Disorder of the Spleen

	NORA		**L1**		**Levin AR**		**RR**		**GPS-EBGM**	
Rank	Concept	Risk	Concept	Risk (OR)	Concept	Risk	Concept	Risk (RR)	Concept	Risk (EBGM)
1	Hemoglobinopathy	0.0702	Hemoglobinopathy	0.8261 (4.75)	Essential hypertension	0.1566	Acute gastric ulcer with hemorrhage AND obstruction	0.9972 (353.8)	Chronic nonalcoholic liver disease	0.8410 (6.29)
2	Primary malignant neoplasm of rectum	0.0146	Hb SS disease	0.7476 (2.96)	Abdominal pain	0.1052	Chronic duodenal ulcer with hemorrhage	0.9972 (353.8)	Hb SS disease	0.8410 (6.29)
3	Ulcer of esophagus	0.0109	Complication of transplanted lung	0.7451 (2.92)	Congestive heart failure	0.0876	Congenital genu recurvatum	0.9972 (353.8)	Hemoglobin SS disease with crisis	0.8410 (6.29)
4	Primary malignant neoplasm of unspecified site	0.0100	Primary malignant neoplasm of unspecified site	0.7150 (2.51)	Anemia	0.0782	Disorder of hair AND/OR hair follicle	0.9972 (353.8)	Hemoglobin SS disease without crisis	0.8410 (6.29)
5	Hereditary elliptocytosis	0.0095	Hemoglobin SS disease with crisis	0.6764 (2.01)	Primary malignant neoplasm of unspecified site	0.0718	Endocrine myopathy	0.9972 (353.8)	Hemoglobinopathy	0.8410 (6.29)

Table 8.9: Hypothyroidism

	NORA		L1		Levin AR		RR		GPS-EBGM	
Rank	Concept	Risk	Concept	Risk (OR)	Concept	Risk	Concept	Risk (RR)	Concept	Risk (EBGM)
1	Complete thyroidectomy	0.5531	Complete thyroidectomy	0.9537 (20.60)	Colonoscopy	0.0394	Arthrodesis	0.9402 (16.73)	Complete thyroidectomy	0.8656 (7.44)
2	Other partial thyroidectomy	0.4037	Other repair of vulva and perineum	0.9220 (11.82)	Upper gastrointestinal endoscopy	0.0253	Arthroplasty	0.9402 (16.73)	Unilateral thyroid lobectomy	0.7996 (4.99)
3	Total thyroid lobectomy, unilateral	0.3503	Other partial thyroidectomy	0.8741 (6.94)	Therapeutic procedure	0.0244	Arthrotomy	0.9402 (16.73)	Other partial thyroidectomy	0.7921 (4.81)
4	Unilateral thyroid lobectomy	0.2911	Unilateral thyroid lobectomy	0.8403 (5.26)	Assessment for prescriptive eye wear	0.0158	Biopsy of back of throat	0.9402 (16.73)	Total thyroid lobectomy, unilateral; with or without isthmusectomy	0.7881 (4.72)
5	Esophagogastroduodenoscopy	0.1725	Enterolysis	0.7472 (2.96)	Ophthalmological services	0.0149	Biopsy of liver	0.9402 (16.73)	Thyroidectomy, total or complete	0.7423 (3.88)

Table 8.10: Mucositis vs Procedures

	NORA		L1		Levin AR		RR		GPS-EBGM	
Rank	Concept	Risk	Concept	Risk (OR)	Concept	Risk	Concept	Risk (RR)	Concept	Risk (EBGM)
1	Professional services for allergen immunotherapy	0.2964	Level 3 outpatient visit ...	0.6424 (1.80)	Level 3 outpatient visit ...	0.1139	Anastomosis of gallbladder to intestine	0.9125 (11.43)	Professional services for allergen immunotherapy	(0.7167) 3.53
2	intercept	0.0774	Level 2 outpatient visit ...	0.5586 (1.27)	Level 2 outpatient visit ...	0.0538	Anesthesia for all closed procedures involving upper two-thirds of femur	0.9125 (11.43)	Rapid desensitization procedure	0.5595 (2.27)
3	Comprehensive audiometry threshold evaluation and speech recognition	0.0680	Level 5 outpatient visit ...	0.5398 (1.17)	Level 4 outpatient visit ...	0.0487	Arthrotomy with biopsy; metacarpophalangeal joint, each	0.9125 (11.43)	Demonstration and/or evaluation of patient use of aerosol generator ...	0.5475 (2.21)
4	Anesthesia for procedures on external, middle, and inner ear including biopsy	0.0561	Level 4 outpatient visit ...	0.5338 (1.15)	Level 5 outpatient visit ...	0.0377	Arthrotomy, glenohumeral joint	0.9125 (11.43)	Anesthesia for procedures on external, middle, and inner ear including biopsy	0.5305 (2.13)
5	Level 3 outpatient visit ...	0.0531	Assessment for prescriptive eye wear using a range of lens powers	0.5202 (1.08)	Interview and evaluation, described as limited	0.0238	Chemotherapy administration, intra-arterial; infusion technique, up to 1 hour	0.9125 (11.43)	Comprehensive audiometry threshold evaluation and speech recognition	0.5192 (2.08)