Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yutong Jin Date

Advances in Causal Inference to Support Vaccine Development and Evaluation

By

Yutong Jin
Doctor of Philosophy

Biostatistics

David Benkeser, Ph.D.
Advisor

Holly Janes, Ph.D.
Committee Member

Robert Lyles, Ph.D.
Committee Member

Hao Wu, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D, MPH
Dean of the James T. Laney School of Graduate Studies

Date

Advances in Causal Inference to Support Vaccine Development and Evaluation

By

Yutong Jin
M.S., Emory University, 2023
MSPH, Emory University, GA, 2018
B.M., Fudan University, 2016

Advisor: David Benkeser, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2023

Abstract

Advances in Causal Inference to Support Vaccine Development and Evaluation
By Yutong Jin

Communicable disease outbreaks continue to present significant challenges to human society. The incidence of various infectious diseases remains alarmingly high globally. Consequently, developing preventive vaccines has become a pivotal objective in mitigating infectious disease burden. This dissertation centers on the creation of statistical methods that support the design and evaluation of potential vaccines.

In the first section, we devise methods to identify key genetic mutations in the HIV envelope protein linked to antibody resistance. This task proves complex due to the high-dimensional and strongly correlated nature of genetic sequence data. We propose a solution using an outcome-adaptive, collaborative targeted minimum loss-based estimation approach combined with random forests, which enjoys significant advantages over existing methods. We apply this approach to the Compile, Analyze and Tally Nab Panels (CATNAP) database to identify amino acid positions causally related to resistance to neutralization by various antibodies.

In the second section, we develop methods for standardized comparisons of immunogenicity across diverse vaccine trials, involving different populations and study designs. To address this, we introduce a causal framework capable of identifying suitable causal estimands and estimators to bridge the immunogenicity of one vaccine from the trial population where it was evaluated to other trial populations. We apply the proposed technique to compare vaccine effectiveness using data from four recent HIV vaccine trials.

In the third section, we create methods to generate standardized versions of causal effects that can be used to compare the impact of vaccines on various outcomes that are measured on different scales. For example, we may wish to compare the difference in immunogenicity between two vaccines in terms of two different immunologic assays. If these assay readouts have substantially different variability, then a comparison of relative vaccine performance across assays can be challenging. To rectify this, we develop a general framework for defining standardized causal effect sizes. We develop nonparametric efficient estimators of these quantities and evaluate the estimators' performance in comprehensive numerical studies.

Advances in Causal Inference to Support Vaccine Development and Evaluation

By

Yutong Jin
M.S., Emory University, 2023
MSPH, Emory University, GA, 2018
B.M., Fudan University, 2016

Advisor: David Benkeser, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2023

Acknowledgments

I would like to express my deepest gratitude and appreciation to my advisor, parents, and husband for their unwavering support, guidance, and love throughout my journey, especially during these challenging pandemic years. COVID-19 has brought about unprecedented difficulties we never anticipated. In the midst of these uncertainty, without their constant encouragement and belief in me, I would not have reached this significant milestone in my life.

To my advisor, Dr. David Benkeser, thank you for your mentorship and dedication to my academic and personal growth. Your patience and insightful feedback have shaped my research and inspired me to step forward reaching new heights. You have always been willing to share your knowledge and expertise with me. Your encouragement and mental support have instilled in me the confidence every time when I felt overwhelmed and stuck in challenging situations. You have always believed in me, even when I doubted myself. Without you, I would not have been able to achieve all milestones I have reached thus far. I am so fortunate to have you as my advisor.

I would also like to extend my heartfelt appreciation to my master supervisor Dr. Zhaohui Qin. Your guidance and support have played a pivotal role in my early academic journey. You were the first person who introduced me to the concept that "there are many ways of problem-solving" and provided invaluable insights into the world of research. I am so grateful for the opportunity to have worked with you and learned from you.

To my parents, words cannot adequately convey my gratitude for your unconditional love, sacrifices, and unwavering belief in my abilities. I am fully aware of the challenges you faced when making the difficult decision to send me to the other side of the globe for my education, especially during these unprecedented times of pandemic. Despite the distance and the uncertainties that came along, you never hesitated to offer your encouragement and support to help me overcome every obstacle. The values, wisdom, and resilience you have instilled in me are immeasurable gifts that I will forever cherish.

To my dear husband, Yunchuan, thank you for being my rock and my biggest cheerleader. Because of you, I never feel lonely and helpless on this land. I am grateful for your presence in my life, and I look forward to sharing many more milestones together with you in the future.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Outbreaks of various communicable diseases remain a major concern in human society. For example, HIV incidence remains stubbornly high - in 2019 36,801 new HIV diagnoses were reported in the United States and at year-end an estimated 1.1 million individuals were living with HIV in the United States [9]. To reduce incidence further, the development of safe and effective preventive vaccines is crucial. In general, vaccines operate by introducing a so-called *antigen* to the host immune system. This antigen could be a fragment of viral genetic material or a protein expressed on a virus. Once the antigen is introduced, the host's immune system is activated, initiating a biological response against the antigen. Subsequently, when the host is later exposed to the real form of the virus, the immune system can recognize it and initiate a range of immune responses aimed at protecting the host from breakthrough infection and disease. There are myriad immune responses generated in response to vaccine antigens. Among these responses, one that is believed to play a vital role in effective vaccines are *antibody responses*. Antibodies are proteins that circulate in the blood and are designed to bind to specific foreign substances, like viruses, thereby *neutralizing* the substance. If a vaccine can successfully stimulate antibodies that are capable

of neutralizing many forms of a particular pathogen (e.g., different strains of a virus), it can be regarded as a promising vaccine for providing robust protection. However, eliciting such broad protection can be challenging for many pathogens including HIV. Thus, the development of broadly effective vaccines remains a considerable challenge.

This dissertation explores several important statistical topics that arise in the vaccine development process. Typically, vaccine development includes three key stages: pre-clinical research, early-phase trials and late-phase trials. The first topic of the dissertation is aimed at challenges that can arise in the pre-clinical stage. We utilize existing observational data sources to identify HIV viruses that may escape neutralization by specific antibodies by estimating the causal effect of an amino acid substitution in the HIV Envelope protein on antibody neutralization. Details are provided in Chapter 2. The second and third topics address practical challenges encountered during clinical trials of vaccines. In the second topic, we propose a framework to identify appropriate causal estimands and estimators that enable objective comparisons of vaccine immunogenicity across trials with different populations and sampling designs. Further information is available in Chapter 3. To facilitate more robust comparisons of immune responses across candidate vaccines, in the third topic, we introduce the idea of a *standardized causal effect size* that may be useful to this end. We also provide nonparametric efficient estimators of the proposed causal estimands. Refer to Chapter 4 for more comprehensive details.

The remainder of this Introduction provides relevant background information for each topic, while specific details of the statistical methods are included in later Sections.

1.2 Targeted Machine Learning for Understanding HIV Resistance to Neutralizing Antibodies

To understand how antibodies are thought to neutralize HIV, it is first important to understand the structure of the virus itself. A simple diagram of HIV is illustrated in Figure

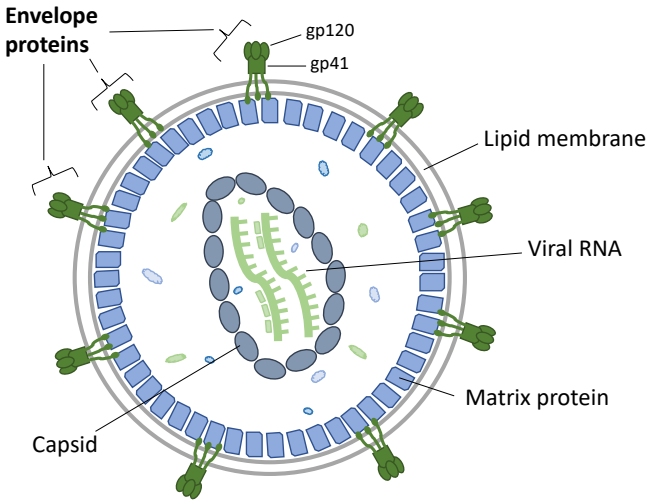

Figure 1.1: Illustration of the structure of HIV. The envelope protein, which includes the gp120 and gp40 sub-proteins, is the putative binding site for antibodies and other immune responses.

1.1. The green spikes protruding from the virus are called *envelope (Env)* proteins and are used by the virus to bind to and subsequently infect host CD4 T-cells. The Env protein is also the putative site for antibody binding. If enough antibodies bind to these Env proteins, they may block the virus from infecting cells, effectively neutralizing the virus. However, HIV can and does evolve specifically to avoid such immune responses and may be able to adapt the geometric shape and/or their physio-chemical makeup of the Env protein to make it more difficult for antibodies to recognize. This presents a challenge for HIV vaccine development – vaccines must induce antibody responses that are capable of binding to and neutralizing a wide variety of Env proteins.

Due to the challenges associated with selecting an antigen capable of inducing such antibodies, an alternative paradigm has emerged for HIV prevention – rather than relying on vaccine antigens to induce neutralizing antibodies, we can instead identify and manufacture such antibodies in a laboratory. This mode of prevention is currently of great interest in the field of HIV prevention. Using sera from HIV-infected humans and/or non-human primates, scientists have identified several so-called broadly neutralizing antibodies

(bnAbs), single antibodies that are capable of neutralizing a wide variety of Env proteins. These antibodies can be manufactured at scale and be given to individuals at risk of HIV acquisition to prevent future HIV infection. Several prevention trials along these lines are being conducted [31], including HVTN-704, a Phase 2b study of a bnAb named *VRC01* [14]. This trial showed modest overall efficacy for preventing HIV, with a multiplicative reduction in risk of infection of about 20%. However, further analysis for this trial revealed a crucial result – individuals given VRC01 were indeed protected from *some* types of HIV, but were susceptible to others. In particular, individuals who were given VRC01 were protected from viruses that were capable of being neutralized by the antibody VRC01, but remained susceptible to infection by HIV viruses that were able to escape VRC01. Investigators demonstrated this effect by sequencing the Env protein from viruses of HIV-infected individuals in the trial. Laboratory experiments were then conducted to measure how well these Env proteins could be neutralized by VRC01. We refer to this outcome measure as the *neutralization sensitivity* of the virus [38]. Combining these data with the clinical data from the randomized trial, researchers estimated prevention efficacy of VRC01 as a function of neutralization sensitivity, demonstrating that receipt of VRC01 was associated with almost no protection from infection with viruses with low neutralization sensitivity, but provided high efficacy against viruses that were highly sensitive to neutralization by VRC01. These results highlighted that bNAbs should be considered a highly promising means of HIV prevention, but, as with vaccines, the key obstacle is how to select antibodies that are capable of neutralizing genetically diverse HIV Env proteins.

Overcoming this challenge will require careful interaction of randomized studies, like HVTN 704 described above, and observational laboratory studies. In the latter, HIV Env sequences isolated from infected participants around the globe are used to measure neutralization sensitivity to various candidate bnAbs. The CATNAP database collates data from these neutralization studies into a central, publicly available repository [45]. However, the number of sequences available for some antibodies is limited and moreover, databases such

as these are not exhaustive in covering all possible configurations of the Env protein that may be observed in the population. Therefore, it is of great interest to develop methods that use observational databases, like CATNAP, to learn the causal effect of mutations in the Env protein on antibody resistance.

Figure 1.2 illustrates the interplay between observational neutralization studies and randomized controlled trials of bnAb therapies. Randomized experiments are generally time-consuming and expensive, and thus we must take care to only advance the most promising bnAbs to randomized trails. One common practice is using neutralization studies to identify promising bnAbs that can be advanced to randomized trials, like HVTN704. If low overall efficacy but significant efficacy against sensitive viruses is found, we may view the trial a partial success in that the bnAb successfully neutralized viruses that it was biologically capable of neutralizing. Then we may consider adding one or several additional bnAbs to create a cocktail of multiple bnAbs. The question then arises as how to select these additional bnAbs. The first valuable source of data are the randomized trials, where we can perform a so-called *sieve analysis* [16].

Sieve analysis quantifies prevention efficacy of the bnAb regimen studied in the randomized trial as a function of characteristics of the Env protein. Because this analysis is based on randomized trial data, it can identify causal effects of Env protein mutations on prevention efficacy under minimal assumptions. However, with more than 800 amino acid (AA) residues in the Env protein, the search space for mutations that may be causally related to antibody resistance is huge. With few observed HIV infections in a randomized trial, sieve analyses that are forced to consider a large number of potential mutations will suffer from low statistical power after accounting for multiple testing. Therefore, it is crucial to utilize observational neutralization data to identify potentially interesting mutations a-priori in order to reduce the impact of multiple testing adjustments.

Causal analysis of neutralization studies can also be used to identify "gaps" in one bnAb – mutations that allow the virus to escape neutralization. If such gaps are identified, we may

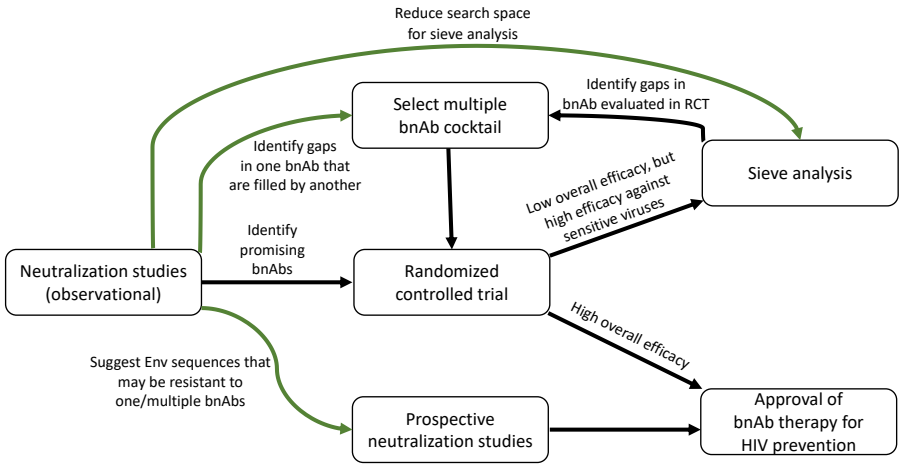

Figure 1.2: Illustration of the interplay between observational neutralization studies and randomized controlled trials of bnAb therapies for prevention. Green arrows illustrate where our proposed method may contribute to the process.

search for additional bnAbs that are robust to these mutations. These results, together with the results of the sieve analysis can be used to select a multi-bnAb cocktail to be evaluated in a future randomized trial.

If and when a successful bnAb or combination of bnAbs are identified, it will also be important to subject such regimen to various "stress tests". That is, we may wish to use observational neutralization studies to suggest mutations that may be particularly difficult for the therapy to neutralize. Neutralization capability of viruses with these mutation patterns can be studied prospectively to suggest whether and to what extent the bnAb therapy will remain effective against new mutations. Such studies may aid in regulatory decision making on the approval of bnAb therapies. For example, during the COVID-19 pandemic, previously approved monoclonal antibody therapies saw their use restricted after the emergence of new variants against which their neutralization capacity was diminished [8]. The above points to a clear need to learn causal relationships between mutations in the Env protein and sensitivity to neutralization by antibodies. Such is the focus of the first chapter of this dissertation and we refer readers to Chapter 2 for further details.

1.3 Comparing HIV Vaccine Immunogenicity across Trials with Different Populations and Study Designs

Vaccine efficacy (VE) is typically quantified as one minus a relative risk, comparing risk of infection or disease under vaccination to risk under a placebo or control vaccine. However, VE is often time-consuming and expensive to assess directly, as randomized trials must wait until a sufficient number of clinical endpoints have been observed. For rare diseases, it can take months and sometimes years to complete such a trial. To avoid this long-lasting and costly process, it is sometimes possible to identify immune responses occurring in response to vaccination that are predictive of VE. Such responses, termed vaccine *correlates of protection* (CoP), may serve as surrogate endpoints in lieu of a formal evaluation of clinical VE. Therefore, in many contexts it is often of interest to study immunological endpoints and compare vaccine immunogenicity across different vaccines. The most common statistical approach to quantifying differences in immune responses across various vaccines is to use a t-test or Wilcoxon Signed-Rank test to test for differences in average immunogenicity or the distribution of immune responses across different vaccines. Sometimes these simple procedures must be extended to account for sampling design, for example by using inverse probability of sampling weighted estimators [4, 11, 15]. However, these approaches are insufficient to provide an objective comparison of vaccine immunogenicity between different vaccines that are evaluated in different studies. These studies may be conducted at different geographic sites and may additionally have discrepant enrollment criteria. This can be problematic for standard approaches to immunogenicity comparison. If there are differences in clinical and/or demographic characteristics across the various trials' populations, and these characteristics also impact immune responses, then confounding is present and simple approaches may yield biased inference regarding differences in vaccine immunogenicity.

Our motivation for studying this problem arises from the field of HIV vaccines. Over

the past decade, there have been many small, and several large studies of preventive HIV vaccines. These trials have been conducted across South East Asia, Sub-Saharan Africa, and in the Americas, each with their own specific set of enrollment criteria. Much of the recent work in HIV vaccine development has been motivated by the results of the RV144 trial, which demonstrated modest but significant efficacy against HIV-1 infection[34]. A key consideration for HIV vaccine development is selection of immunogens for vaccines. Immunogens are molecules that are capable of eliciting a host immune response and must be carefully selected in order for vaccines to generate protective immune responses. The vaccine studied in the RV144 trial consisted of two immunogens, with one based on regionally circulating strains in Thailand. While the study demonstrated modest preventive efficacy with an estimated 31.2% reduction of the cumulative incidence of HIV-1 infection, the vaccine was not licensed for broad use. Nevertheless, the result was encouraging and prompted several smaller follow-up trials including the HVTN097 trial, which was designed to evaluate immunogenicity of the same vaccine regimen in South Africa [18]. The results indicated that response rates and magnitudes of putatively protective immune responses in South Africa were at least as good as those observed in Thailand providing support for continuing with this vaccine platform for research in Sub-Saharan Africa. A subsequent study, HVTN100, investigated a form of the RV144 vaccine but updated with immunogens based on HIV-1 subtypes prevalent in South Africa. This trial also concluded that the South African-adapted vaccine successfully met all pre-specified immunological criteria [5]. Based on these results, a larger phase IIb/III randomized efficacy trial, HVTN702, was conducted to characterize the clinical efficacy of the adapted vaccine for preventing HIV-1 infection in South Africa. Unfortunately, this study was halted during an interim review according to a pre-specified futility criteria, demonstrating no evidence of clinical efficacy to prevent HIV-1 infection [19]. To help identify potential explanations for the lack of efficacy in HVTN702 and to determine the next directions for the HIV vaccine field, it is critical to examine possible difference in immunogenicity between the vaccines

used across RV144, HVTN097, HVTN100, and HVTN702. However, this may be challenging due to the fact that the vaccines were evaluated in different study populations and using trials with a variety of designs.

The need for a comparison of vaccine-induced immune responses across vaccines that are evaluated in trials is not unique to HIV vaccines. Indeed, this is a common and important problem in many domains of vaccine research, including in recent evaluations of preventive COVID-19 vaccines. Several large randomized studies were launched to evaluate efficacy of COVID-19 vaccines. However, the design and enrollment characteristics of these trials differed considerably. For example, the Coronavirus Efficacy (COVE) study evaluated the Moderna mRNA1273 vaccine and used a case-cohort design to sample immune responses in participants, while the Phase 3 study of the the Pfizer/bioNtech BNT162b2 vaccine used random sampling. Moreover, participants in the COVE study were considerably older and more racially diverse than those enrolled in the Pfizer/bioNtech study. Typically, vaccines elicit weaker immune responses in older adults, such that directly comparing the immunogenicity of two vaccines based on these studies' results is challenging. Similar issues arise in studies of dengue vaccines, where past dengue exposure may be a key modifier of the immunogenicity and efficacy of the vaccines[33]. Thus, comparing immunogenicity of the vaccine in populations with differing distributions of exposure histories is a key challenge.

The above scientific context highlights a clear need for understanding causal relationship between a vaccine regimen and the immunogenicity in a particular population. Such information can guide the design of new vaccines, as well as for prioritize current vaccines for further research. This is the aim of the second topic of this dissertation, with details included in Chapter 3.

1.4 Standardized Causal Effect Sizes in Biomedical Research

In many quantitative disciplines, there is a growing interest in estimating causal effects of interventions using observational data [24, 27, 1]. Causal effects are typically formulated as a comparison between potential outcomes [36]. In this framework, we conceptualize that each individual would have experienced different outcomes had they received different interventions. The difference between these outcomes is then summarized to represent the causal impact of certain intervention. A common approach to quantifying causal effects is to consider the *average* potential outcomes under each intervention. The comparison of these averages across different interventions is commonly referred to as the *average treatment effect* (ATE).

In many fields of research it is often of interest to compare the magnitude of intervention effects across *multiple outcomes*. For example, in vaccine research, it is common to study the impact of different vaccine formulations on several distinct immune responses, including antibody responses and T-cell responses. In such settings, it is natural to inquire as to how the relative impact of one intervention versus another differs across outcomes. If the distribution of outcomes differs considerably, comparison of the ATE across outcomes may not always be informative to this end. Returning to the vaccine example, consider a T-cell assay that typically outputs values approximately uniformly over the range 1-10, and an antibody assay that typically outputs values approximately uniformly over the range $1\text{-}10^3$. A comparison of two vaccines may yield an ATE of 5 for both immune responses. However, we expect that this ATE represents a far more dramatic impact on the T-cell response than it does on the antibody response. Thus, reporting the ATE alone does not always appropriately elucidate the impact of interventions when variability of outcomes is ignored.

This discussion underscores the need to develop additional causal estimands beyond

ATE that appropriately reflect the impact of interventions in a maximally interpretable and comparable way. This idea has existed in the associational literature for sometime, where it is widely recognized that associations should somehow account for the inherent variability of an outcome in order to accurately reflect the strength of an association. Such estimands are often referred to as *standardized effect sizes*. These measures typically consist of a ratio of some measure of association in the numerator versus some measure of outcome variability in the denominator. A wide array of estimators has been proposed for estimating standardized effect sizes. Perhaps the most commonly used is Cohen's d, which equals the difference of in average responses between two populations divided by some form of standard deviation (SD) of the response [12]. Exactly which SD should be used is the matter of some debate, with various options proposed for settings where the two populations have equal/unequal outcome variability and where outcomes are/are not normally distributed. Glass's delta proposes to divide by the SD of the control group [17], while Hedge's g focuses on the issue of uneven population sizes [22]. These three measures may be most useful when outcomes are normally distributed and the assumption of equal variances holds [20].

In spite of the name, standardized effect sizes are rarely dealt with in an explicit causal framework and there has been little exploration of whether and how such measures could be adopted to an explicit causal framework. In the third project of this dissertation, we propose an explicit adaptation of these measures to a causal setting and discuss relevant issues in defining appropriate standardized causal effect sizes. For details, see Chapter 4.

Chapter 2

Identifying HIV sequences that escape antibody neutralization using random forests and collaborative targeted learning

2.1 Introduction

In this chapter, we aim to identify complementary bnAbs that target the residues where some existing antibodies show modest protection. Additionally, we provide a useful hypothesis test that guides downstream sieve analysis by reducing the search space. To achieve these goals, we thus discuss whether and how the causal effect of a mutation at a single AA residue in the Env protein on antibody sensitivity can be identified and estimated from observational neutralization data. We discuss the plausibility of the assumptions in the context of known HIV biology and the data that are typically available in observational neutralization studies. We also propose an approach for dealing with the challenges associated with practical violations of the positivity assumption. In particular, we develop a random forest-based, outcome-adaptive, collaborative targeted minimum loss-based estimation (CTMLE) approach. Our estimation and inference is built on two models. The first is a model of the probability of sensitivity as a function of Env AA residues and other measured viral characteristics, or the so-called outcome regression (OR). The second is a model of the distribution of AAs at a particular residue as a function of other sequence features, or the so-called generalized propensity score (GPS). For both models, in this work we focus on the random forest algorithm [7], which has been shown to predict well in this setting [30].

A typical CTMLE implementation uses a greedy forward selection algorithm that sequentially seeks best next candidate feature to be included in each iteration [44]. However, this approach yields heavy computational burden and it motivates us to pursue a method that is able to reduce time complexity. Our work is similar, yet distinct from the partial-correlation-based pre-ordering of covariates suggested by Ju et al. [29]. Rather than pre-ordering covariates, we instead adopt an explicitly *outcome-adaptive approach* along the lines of [39]. That is, we advance the most relevant features identified in the outcome regression for inclusion in the model for the generalized propensity score. We show that this stabilizes inference relative to procedures that include all features in the generalized

propensity model. The modification results in significant time savings relative to [29], which is particularly important given the high-throughput nature of the scientific context. Whereas Ju et al. [29] were focused on methods for estimating the effect of a single treatment, our analysis requires estimation of an effect for each of several hundreds of AA residues. We demonstrate that our tests based on our CTMLE estimates appropriately controls type I error in realistic sample sizes, even in the presence of highly correlated residues, while maintaining power to detect effects of AA mutations in Env on antibody sensitivity.

2.2 Methods

2.2.1 Counterfactual antibody resistance probability

We assume that we have access to a database like the CATNAP database, described in Chapter 1. An example illustration of one such data set is given in Table 2.1. The data consist of n HIV Env sequences. For each sequence, we have a measure of sensitivity to an antibody of interest (e.g., 1 = can effectively be neutralized by the antibody, 0 = cannot) and basic information about the origin of the sequence. The remaining columns represent the amino acids that make up the Env proteins of the sequence. Each amino acid (AA) position, or *residue*, of the Env protein can assume one of 22 different values – one of the 20 amino acids that build proteins, a *frameshift* mutation (indicating an insertion or deletion of nucleotide bases), or a *gap* (an artefact of the sequencing technology to maintain alignment with a referent HIV sequence). For our purposes, it suffices to think of each AA residue as a categorical variable that theoretically could assume up to 22 different values, while in practice we generally observe between two and four unique AAs at each residue. We treat the n sequences as independent, which may be reasonable because (i) almost all sequences in CATNAP are isolated from different individuals and (ii) HIV replicates extremely rapidly, such that any two isolated sequences, even if close in geographic proximity, are likely to be distant ancestors of one another.

Sequence ID	Sensitive (Y)	Origin (W_0)	AA_1 (W_1)	AA_2 (W_2)	...	AA_{856} (W_J)
1	1	Africa	N	R	...	L
2	0	Europe	N	K	...	Q
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	1	N. America	S	R	...	L

Table 2.1: Example data set illustrating structure of CATNAP data for a given antibody. Sensitive is a binary read out from the assay indicating whether the sequence was effectively neutralized by the antibody. AA_j = amino acid at residue j of the Env protein.

We denote the sensitivity outcome by Y, where $Y = 1$ indicates that the virus is effectively neutralized by the antibody, and $Y = 0$ indicates the virus is not effectively neutralized by the antibody. We denote the origin of the virus together with the vector of AA information on the Env sequence as $\boldsymbol{W}$. For simplification, basic information about the origin is encoded as W_0. Below, we consider methods for analyzing the causal impact of mutations at each of the J AA residues in turn. It is therefore convenient to introduce the notation $\boldsymbol{W}_{-j}$ to denote the origins and the vector of all Env characteristics with the j-th residue removed, where $j = 1, \ldots, J$. Thus, the data that we use for the analysis of the j-th residue can be considered n copies of the triplets $\boldsymbol{O}_j = (W_j, \boldsymbol{W}_{-j}, Y)$.

For each AA residue, we define a counterfactual resistance outcome $Y(W_j = w)$, which is the binary resistance indicator that would be observed under a hypothetical intervention that fixes the AA at residue j to $w \in \mathscr{W}_j$, the set of possible AAs at residue j that could be observed in the entire population of virus Env sequences. We also define $p_j(w) = P[Y(W_j = w) = 1]$ as the proportion of virus Env sequences in this counterfactual world that would be sensitive to neutralization by the bnAb of interest.

To achieve the goals outlined in Figure 1.2, we may be interested in estimation and inference about $p_j(w)$. For example, to identify gaps in one bnAb, we would be interested in identifying (j, w) combinations for which $p_j(w)$ is low. To complement such a bnAb in a multiple bnAb cocktail, we would look for other antibodies such that the same counterfactual sensitivity probability is high. On the other hand, if the scientific goal of the analysis of the observational neutralization data is to reduce the search space for a sieve analysis,

we may instead wish to test the null hypothesis that an AA substitution at residue j has no impact on average bnAb resistance, that is, $H_0 : p_j(w)$ is constant in w. AA residues for which this null hypothesis is rejected may be advanced for further consideration in a sieve analysis using data from a randomized trial. Regardless of the ultimate scientific goal, the problem of identification and estimation of counterfactual sensitivity probabilities is an important problem.

2.2.2 Causal Identification

We require two main conditions for identifiability of $p_j(w)$: (i) conditional exchangeability and (ii) positivity. If these conditions hold then $p_j(w)$ is identified by the G-functional $p_j(w) = E[E(Y \mid W_j = w, \boldsymbol{W}_{-j})]$. We discuss these conditions in detail below and conclude with a discussion of their plausibility in the present context.

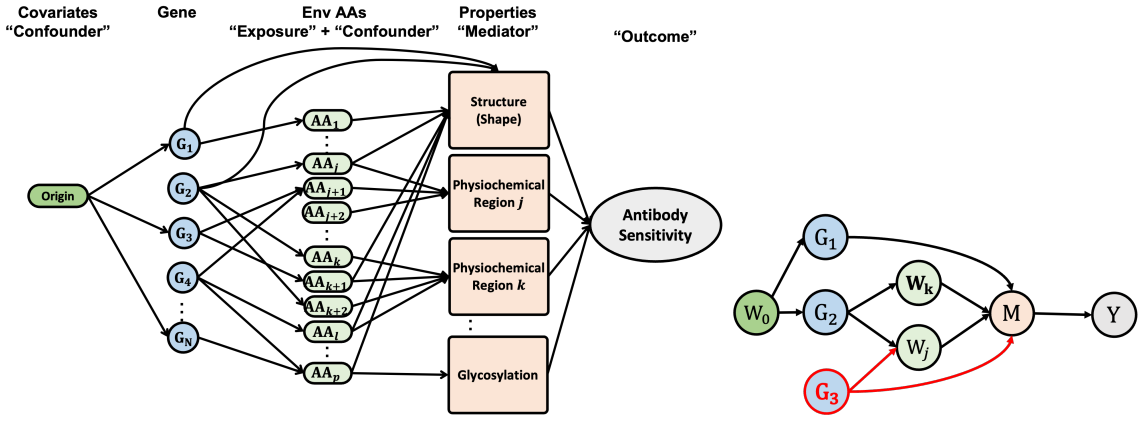

Figure 2.1: **DAG for HIV resistance.** The DAG on the left shows a pipeline of gene expression, which defines how gene regulates downstream activities and further affect the antibody sensitivity through Envelope amino acids of interest.

Exchangeability

A DAG is displayed in Figure 2.1. On the left, we display a DAG that is useful for discussing the biology of antibody neutralization. On the right, we display a DAG that is

useful for describing backdoor pathways and types of unmeasured confounders that may prove problematic for our proposal.

Beginning with the DAG on the left, the nodes labeled G represent genes on the viral genome. Certain genes on the genome correspond with AA residues on the Env protein (nodes labeled AA). These AA configurations result in certain physiochemical properties of the Env protein. The DAG lists several Env characteristic that are putatively important for antibody binding and neutralization. These characteristics act as *mediators* of AA configuration on antibody sensitivity. The structure (or shape) of the Env protein is determined by the AAs present; this in turn determines which portions of the protein are "exposed" to antibodies. There are other physiochemical properties that are also impacted by the Env protein AA sequence that could impact affinity of antibodies for binding the protein. One particular process that may be important for HIV infectivity and antibody sensitivity is the presence of glycosylated regions of the Env protein. Such regions can play a key role in binding, both of the virus to the host CD4 T-cells, as well as antibodies to the virus. Therefore these regions are expected to play an especially important causal role in neutralization of viruses via antibodies.

A simplified DAG is shown on the right with nodes collapsed by the role they play in opening/closing backdoor paths between the outcome and AA residues in the protein. As mentioned above, our approach considers each AA in turn. So for a particular residue j, W_j is acting as the "*exposure*" of interest, while all other AA residues are acting as potential "*confounders*" $\boldsymbol{W_k}$. The genes are unmeasured and play the role of potential *unmeasured confounders*, while the origin of the virus can be considered another *potential confounder* W_0. The mediators in this problem, here collapsed to a single node M, are the aforementioned structural or physiochemical properties of the Env protein. The DAG on the left suffices to discuss the biological plausibility of causal identification in this problem. In particular, there are three types of pathways arising to be relevant to identification. The first are pathways of the form $W_j \leftarrow G_2 \rightarrow \boldsymbol{W_k} \rightarrow M \rightarrow Y$. In such pathways, there is

a gene responsible for coding both the AA residue of interest, as well as other residues that are associated with mediators of antibody sensitivity. In this case, controlling for $\boldsymbol{W_k}$ suffices to block the pathway. The second type of pathway is one of the form $W_j \leftarrow G_2 \leftarrow W_0 \rightarrow G_1 \rightarrow M \rightarrow Y$. Here, the geographic region (W_0) is associated with two genes that code separate residues on the Env protein. This pathway can be blocked by conditioning on W_0. Finally, there are pathways of the form $W_j \leftarrow G_3 \rightarrow M \rightarrow Y$. Here, there is a gene that codes the AA residue of interest, but is also associated with some other mediating pathway of antibody sensitivity. For example, it is possible that having more Env proteins expressed reduces antibody sensitivity of viruses. If a particular gene was associated with the number of Env proteins that are expressed and also associated with the AA residue of interest, then there would be no way to block this pathway, since the gene information is unavailable. Thus, in order to formally establish causal identification, we would need to assume that no such genes exist. If not, then the DAG implies exchangeability, that is $Y(W_j = w) \perp W_j \mid \boldsymbol{W}_{-j}$.

Positivity

Causal identification would additionally require that $P[P(W_j = w \mid \boldsymbol{W}_{-j}) > 0] = 1$ for all $w \in \mathscr{W}_j$. That is, we require that there be a positive probability of observing AA w at position j given the AA present at other positions and the country of origin. This assumption could be satisfied by a careful selection of $\mathscr{W}_j$. In practice, we will establish $\mathscr{W}_j$ empirically by looking at the observed support of each W_j. However, this alone is insufficient to ensure the positivity condition holds. We must additionally assume that any observed AA substitution is plausible, *regardless* of the other amino acids present in the Env sequence. This assumption may be plausible for some residues, but implausible for others depending on the physio-chemical properties of the amino acids in question. Thus, in practice it may be possible to have structural violations of this assumption. We also expect practical violations of this assumption, with some combinations of amino acid configurations only rarely

observed in the source population.

Plausibility of assumptions

The above discussion highlights the tenuous nature of causal interpretations in the context of antibody sensitivity. We generally do not yet possess a strong enough understanding of HIV biology to fully justify the exchangeability condition, and it is also possible that structural violations of the positivity assumption may be present. If the scientific context of an analysis demands a confirmatory causal conclusion to be drawn, then we may require additional causal sensitivity analysis to determine the robustness of the estimated effects to the presence of unmeasured confounding pathways. However, we argue that a causally-oriented analysis may still provide a useful framework for hypothesis generation and exploratory analysis in the context of the research process described in Figure 1.2. In this case, we may prefer to interpret the G-functional identification parameter more cautiously, and summarize the "variable importance" measures of particular AA residues, as in [40]. In this case, follow up experiments should be recommended to validate the findings.

2.2.3 Motivation for novel method

There are many methods available that could in principle be applied in a non-causal context in this setting. For example, we could first conduct an analysis using LASSO on half of samples. The best value of the tuning parameter λ can be determined via cross-validation to arrive at a final model. The AA residues included in the final model can be included in a logistic regression on the withheld half of samples for formal inference. Another approach would be to use random forest "variable importance" measures, for which we may obtain p-values using a fast variable importance test [28]. However, we argue that methods that have a causal interpretation, even if under limited circumstances, are appealing in this setting since causal inference is fundamentally at the core of the scientific question. Thus, we are motivated to explore causal inference-inspired methodology for estimating the G-

functional $E[E(Y \mid W_j = w, \boldsymbol{W}_{-j})]$ and associated significance tests for AA residues.

There are two related challenges associated with causal inference methods in this context: (i) the dimensionality of the covariates and (ii) practical violations of the positivity assumption. The dimensionality of the covariates presents challenges to estimation. The discussion of exchangeability above highlights the need to potentially adjust for a high-dimensional $\boldsymbol{W}_j$, with a limited number of sequences available for a given analysis. Thus, we may be motivated to consider flexible machine learning algorithms that are built specifically for modelling high-dimensional covariates. It would be natural to couple these modelling strategies with doubly-robust methods for inference, as these methods typically allow for regular, parametric-rate inference, under certain statistical assumptions. However, these methods can be susceptible to erratic behavior in the presence of practical positivity violations. One approach to improving behavior of doubly robust estimators in the presence of positivity violations involves a careful pre-selection of adjustment covariates, informed by background knowledge [32]. However, this would appear difficult in the present setting due to imperfect understanding of the relationships between AA residues on the Env protein and their relationship with antibody neutralization. Moreover, for a method to be successfully employed to scan the entire Env protein for significant residues, it would need to be a high throughput method that is capable of delivering stable inference on hundreds of AA residues. An analysis that involves deliberate pre-selection of covariates AA residue-by-residue is infeasible. Therefore, we are motivated to develop an automated procedure for covariate selection in this context. In the subsequent sections, we propose a solution for this problem using collaborative TMLE (CTMLE).

2.2.4 TMLE

Our CTMLE builds on targeted minimum loss-based estimators (TMLE) of the average treatment effects [42]. TMLEs are constructed using estimates of two key quantities: the outcome regression and the generalized propensity score. The OR is defined as the condi-

tional mean of the outcome given the treatment and confounding factors, which we denote by $\bar{Q}(W_j, \boldsymbol{W}_{-j}) = P(Y = 1 \mid W_j, \boldsymbol{W}_{-j})$. The GPS is the conditional distribution of AAs at residue j given all other AA residues and sequence information, which we denote by $g_j(w \mid \boldsymbol{W}_{-j}) = P(W_j = w \mid \boldsymbol{W}_{-j})$ for $w \in \mathscr{W}_j$. A TMLE for estimating $E[Y(W_j = w)]$ for a particular j and w can be constructed in two steps. In step one, initial estimator $\bar{Q}_n$ of the OR can be obtained using any learning technique for classification of a binary outcome. Similarly, the estimator $g_{n,j}$ of GPS could be estimated using any multi-class classification technique. In step two, a single iteration of boosting is used to update the initial OR. For a given OR estimate, $\bar{Q}$, we can compute its empirical risk,

$$\mathrm{L}_{n,j}(\bar{Q}) = \frac{1}{n}\sum_{i=1}^{n} - \left[Y_i \log\{\bar{Q}(W_{j,i}, \boldsymbol{W}_{-j,i})\} + (1 - Y_i)\log\{1 - \bar{Q}(W_{j,i}, \boldsymbol{W}_{-j,i})\}\right] . \quad (2.1)$$

Next, we update $\bar{Q}_n$ by minimizing empirical loss (2.1) along a particular univariate regression model indexed by $\bar{Q}_n$. For each w, we define the logistic regression model

$$\bar{Q}_{n,\varepsilon(w)}(W_j, \boldsymbol{W}_{-j}) = \mathrm{expit}\left[\mathrm{logit}\{\bar{Q}_n(W_j, \boldsymbol{W}_{-j})\} + \varepsilon(w)\frac{I(W_j = w)}{g_{n,j}(w \mid \boldsymbol{W}_{-j})}\right] , \ \varepsilon(w) \in \mathbb{R} .$$

We can obtain a maximum likelihood estimate $\varepsilon_n(w) = \arg\min_{\varepsilon(w)} \mathrm{L}_{n,j}(\bar{Q}_{n,\varepsilon(w)})$ and define the updated OR estimate

$$\bar{Q}^*_n(w, \boldsymbol{W}_{-j}) = \mathrm{expit}\left[\mathrm{logit}\{\bar{Q}_n(w, \boldsymbol{W}_{-j})\} + \frac{\varepsilon_n(w)}{g_{n,j}(w \mid \boldsymbol{W}_{-j})}\right] . \quad (2.2)$$

The final estimate is computed as

$$\hat{p}_j(w) = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}^*_n(w, \boldsymbol{W}_{-j,i}) .$$

The key idea motivating the second step of the TMLE is that the bias of the revised OR estimator $\bar{Q}^*_n$ is generally smaller than that of $\bar{Q}_n$ with respect to estimating $p_j(w)$.

This procedure can be repeated for each $w \in \mathscr{W}_j$. With a minor abuse of notation, we define the implied updated estimate of $E\{Y \mid W_j, \boldsymbol{W}_{-j}\}$ as

$$\bar{Q}_n^*(W_j, \boldsymbol{W}_{-j}) = \text{expit}\left[\text{logit}\{\bar{Q}_n(W_j, \boldsymbol{W}_{-j})\} + \sum_{w \in \mathscr{W}_j} \varepsilon_n(w) \frac{I(W_j = w)}{g_{n,j}(w \mid \boldsymbol{W}_{-j})}\right] .$$

Recalling our null hypothesis of interest that $E[Y(W_j = w)]$ is constant in w, a TMLE-based test based on these estimates can be derived as follows. For $w \in \mathscr{W}_j$, we define

$$D_{j,i}(w) = \frac{I(W_{j,i} = w)}{g_{n,j}(w \mid \boldsymbol{W}_{-j,i})}\{Y - \bar{Q}_n(W_{j,i}, \boldsymbol{W}_{-j,i})\} + \bar{Q}_n(W_{j,i}, \boldsymbol{W}_{-j,i}) - \hat{p}_j(w) ,$$

and let $\boldsymbol{D}_{j,i}$ be a $|\mathscr{W}_j|$-length row vector consisting of $D_{j,i}(w)$ for each $w \in \mathscr{W}_j$. Let $\boldsymbol{D}_j$ be a $n \times |\mathscr{W}_j|$ matrix formed by stacking the row vectors $\boldsymbol{D}_{j,i}, i = 1, \ldots, n$. The estimates $\hat{\boldsymbol{p}}_j = \{\hat{p}_j(w) : w \in \mathscr{W}_j\}$ of $\boldsymbol{p}_j = \{p_j(w) : w \in \mathscr{W}_j\}$ have asymptotic covariance that is consistently estimated by $\hat{\Sigma} = n^{-1}\boldsymbol{D}_j^\top \boldsymbol{D}_j$.

We propose to test H_0 using a $\binom{|\mathscr{W}_j|}{2} - 1$ degree-of-freedom Wald test. Specifically, we define a contrast matrix $\boldsymbol{A}$, where each row defines a contrast between $\hat{p}_j(w)$ and $\hat{p}_j(w')$ for $w, w' \in \mathscr{W}_j$. For example, if $|\mathscr{W}_j| = 4$, then

$$\boldsymbol{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} ,$$

and our null hypothesis can be written as $H_0 : \boldsymbol{A}\boldsymbol{p}_j = 0$.

If instead inference on a single $p_j(w)$ is desired, the standard error of $\hat{p}_j(w)$ is given by the square root of the (j, j) element of $\hat{\Sigma}$. Assuming n is large enough, a two-sided Wald-style hypothesis test that rejects $H_0 : \hat{p}_j(w) = \beta$ whenever $|\hat{p}_j(w) - \beta|/\hat{\sigma}_n$ is greater than the $1 - \alpha/2$ quantile of a standard normal distribution will have type I error no larger than α.

2.2.5 CTMLE

A challenge with utilizing TMLE in the present setting is that the estimated GPS may be extremely small for some virus sequences due to high correlations between residues in the Env protein. This can lead to erratic estimates $\varepsilon_n(w)$ of $\varepsilon(w)$, which can degrade the OR estimate $\bar{Q}_n^*$ in (2.2) considerably. The resulting ATE estimate could be highly biased, with correspondingly inflated type I errors and poor confidence interval coverage. There are fixes available [32]; however, the approaches require manual manipulation from the analysts (e.g., to identify overlapping covariate regions and redefine the causal parameter accordingly). Unfortunately, in our motivating example we wish to consider hundreds of AA residues. This motivates the pursuit of a more high throughput analysis approach that maintains stability even in rather extreme scenarios.

Recently, data-driven PS model-building strategies have emerged [39, 29]. A key insight of these approaches is that PS models should adjust only for variables that are related to the outcome. Thus, estimated ORs can be used to inform variable inclusion in propensity score models. By appropriately screening so-called instrumental variables (those that impact only the GPS but not the OR), we may generate less extreme estimates of the GPS and thereby attain more stable behavior of TMLE.

In particular, we propose using variable importance measures from an OR model to select variables to include in the GPS model. We focus on random forest models explicitly, for estimation of both the OR and GPS, though the methods theoretically extend to any machine learning framework that has associated variable importance measures. We focus on random forests in particular for two reasons: (i) robust and fast software implementations with variable importance measures are readily available; (ii) past work has shown that random forests perform extremely well in predicting antibody neutralization, even when compared with other state-of-the-art approaches like deep learning [30, 10].

Our approach entails first using random forests to estimate the OR. During the random forest construction, variable importance for each AA residue is summarized by the mean

decrease in Gini index. The Gini index quantifies how much each feature contributes to the decline of node impurities on average. We select a fixed number of the top-ranked features to include into GPS model. The OR and GPS are then combined into estimates of $E[Y(W_j = w)]$ for each j and w and a test of the significance of each AA residue. We propose a CTMLE algorithm for selecting the optimal number of features to include in the GPS model.

Details of CTMLE

A CTMLE-based estimate of $p_j(w)$ can be obtained in a sequential algorithm as follows.

- Fit OR model using random forests to get an initial estimator $\bar{Q}_n^{(1)}$. All covariates should be included in this model. The covariates are ranked by their feature importance.

- Propose K potential values, $r_1, \ldots, r_K$, of the number of covariates to be included in the GPS model. We assume $r_K = J$, though that need not be the case. A sequence of GPS estimators can be constructed as $g_{n,j}^{(1)}, \ldots, g_{n,j}^{(K)}$. The k-th estimator in this sequence is an estimate of the conditional distribution of AAs at residue j given the r_k top-ranked features in the feature importance list specified in the previous step. If residue j is included in the r_k top-ranked features, it will be excluded and only $(k-1)$ features will be used in GPS estimation.

- If $k = 1$, an initial triplet is built as $\left(g_{n,j}^{(1)}, \bar{Q}_{n,j}^{(1)}, \bar{Q}_{n,j}^{*,(1)}\right)$, where $\bar{Q}_{n,j}^{*,(1)}$ is a fluctuation of $\bar{Q}_{n,j}^{(1)} = \bar{Q}_n^{(1)}$ using TMLE algorithm presented in Section 2.2.4. For $k = 2, \ldots, K$:

  1. Assign $\bar{Q}_{n,j}^{(k)} = \bar{Q}_{n,j}^{(k-1)}$.
  2. Obtain a corresponding $\bar{Q}_{n,j}^{*,(k)}$ by performing similar TMLE steps using $\bar{Q}_{n,j}^{(k)}$ and $g_{n,j}^{(k)}$.
  3. Evaluate $\mathrm{L}_{n,j}(\bar{Q}_{n,j}^{*,(k)})$ as defined in equation (2.1). If $\mathrm{L}_{n,j}(\bar{Q}_{n,j}^{*,(k)}) \leq \mathrm{L}_{n,j}(\bar{Q}_{n,j}^{*,(k-1)})$, then set the k-th triplet to $\left(g_{n,j}^{(k)}, \bar{Q}_{n,j}^{(k)}, \bar{Q}_{n,j}^{*,(k)}\right)$; otherwise, set $\bar{Q}_{n,j}^{(k)} = \bar{Q}_{n,j}^{*,(k-1)}$ and

repeat steps 2-3. In this case, by the construction of the TMLE, it must be true that $\mathrm{L}_n(\bar{Q}^{*,(k)}_{n,j}) \leq \mathrm{L}_n(\bar{Q}^{*,(k-1)}_{n,j})$, since the TMLE $\bar{Q}^{*,(k)}_{n,j}$ uses $\bar{Q}^{*,(k-1)}_{n,j}$ as initial estimate.

Once K triplets have been derived using the full data, the above procedure is repeated in V training samples to get K training-sample specific triplets, denoted for the v-th training sample by $\left(g^{(k)}_{n,j,v}, \bar{Q}^{(k)}_{n,j,v}, \bar{Q}^{*,(k)}_{n,j,v}\right)$. Let $S_v \subseteq \{1,\ldots,n\}$ denote the indices of observations in the v-th validation sample. For $v = 1,\ldots,V$ and $k = 1,\ldots,K$, we compute

$$\mathrm{L}_{n,v}(\bar{Q}^{*,(k)}_{n,j,v}) = \frac{1}{|S_v|} \sum_{i \in S_v} - \left[Y_i \log\{\bar{Q}^{*,(k)}_{n,j,v}(W_{j,i}, \boldsymbol{W}_{-j,i})\} + (1 - Y_i) \log\{1 - \bar{Q}^{*,(k)}_{n,j,v}(W_{j,i}, \boldsymbol{W}_{-j,i})\} \right],$$

and select the choice of k that minimizes the cross-validated risk, $k_n = \arg\min_k \sum_{v=1}^{V} \mathrm{L}_{n,v}(\bar{Q}^{*,(k)}_{n,j,v})$. The final CTMLE estimate is

$$\hat{p}_{j,\mathrm{CTMLE}}(w) = \frac{1}{n} \sum_{i=1}^{n} \bar{Q}^{*,(k_n)}_{n,j}(w, \boldsymbol{W}_{-j,i}).$$

Asymptotic properties of this estimator follow from general results pertaining to CTMLE [44].

2.3 Simulation study

2.3.1 Design

We conducted Monte-Carlo simulation studies to compare the performance of the proposed CTMLE to several implementations of TMLE. The first, was a standard implementation of TMLE, where all features were included in the GPS model, while the other TMLEs used reduced numbers of features (either 5, 10, 50, or 100) based on their estimated importance in the OR model. For all estimators, GPS estimates were truncated at 0.01. We considered two sample sizes $n \in \{500, 1000\}$, which are similar to the sample sizes of the CATNAP

data. For each sample size, 1000 data sets were generated as follows. To represent a sequence of AA residues, we simulated 200 moderately correlated categorical variables, $W_1, W_2, \ldots, W_{200}$, each containing 4 levels, $W_j \in \{1,2,3,4\}$. To generate a realization of $\boldsymbol{W} = (W_1, \ldots, W_{200})^\top$, we drew a random variable $\boldsymbol{X} = (X_1, \ldots, X_{200})^\top$ from a multivariate normal distribution $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{\mu}_0 = (0,0,0,0)^\top$ and an autoregressive-1 covariance structure, $\boldsymbol{\Sigma_0} = \sigma^2\boldsymbol{\Sigma}(\rho)$, where $\boldsymbol{\Sigma}(\rho)$ is a matrix with (j,k)-th entry equal to $\rho^{|j-k|}$. In our example, σ and ρ were fixed at 0.9 and 0.75, indicating a moderate correlation between adjacent features. Next, we set $W_j = 1$ if $X_j < q_{0.25}$, $W_j = 2$ if $q_{0.25} < X_j < q_{0.50}$, $W_j = 3$ if $q_{0.50} < X_j < q_{0.75}$, and $W_j = 4$ if $X_j > q_{0.75}$, where q_p is the p-th quantile of a standard normal random variable. Five true signals were randomly selected among all features (W_j : $j \in \mathscr{J}$, where $\mathscr{J} = \{37, 87, 94, 135, 151\}$). Given $\boldsymbol{W} = \boldsymbol{W}_i$, the outcome Y_i was generated from a Bernoulli distribution with success probability $= \text{expit}\{\sum_{j\in\mathscr{J}}\sum_{\ell=1}^{4}\beta_{jl}I(W_{j,i} = \ell)\}$ (Table 2.2).

	β_{j1}	β_{j2}	β_{j3}	β_{j4}
W_{37}	0.160	-0.321	-0.492	0.214
W_{87}	0.181	0.521	-0.612	0.321
W_{94}	0.104	0.414	-0.789	-0.117
W_{135}	0.178	0.350	-0.453	-0.433
W_{151}	0.072	0.311	0.638	-0.320

Table 2.2: True coefficients (β) used in simulation study

We focused on evaluating operating characteristics for three specific AA residues: (i) position 10, a position unrelated to the outcome and essentially uncorrelated with any true signals; (ii) position 85, a position unrelated to the outcome, but highly correlated a true signal; and (iii) position 87, a position that is truly related to the outcome ($p_{87}(1) = 0.514, p_{87}(2) = 0.588, p_{87}(3) = 0.342, p_{87}(4) = 0.545$). For each of these sites, the C/TMLE point estimates were evaluated by their bias and mean-squared error (MSE) and we compared the power of the Wald test for rejecting the null hypothesis of no relationship between AAs and outcome.

2.3.2 Results

We found that the bias and MSE of the TMLE estimators tended to increase as more variables were included in the GPS model (Figure 1, left/middle column), with the largest bias seen for the true signal (position 87, blue plus). The CTMLE had considerably reduced bias and MSE relative to the standard TMLE that included all 200 features in the GPS model. Moreover, in the two null scenarios (positions 10 and 85), the standard TMLE had drastically inflated type I error (Figure 2.2, right column). For example, the tests incorrectly rejected the null hypothesis more than 60% of the time when $n = 1000$. On the other hand, CTMLE-based tests appropriately rejected the null hypothesis about 5% of the time for the null positions. For the position exhibiting a true signal (position 87), CTMLE rejected the null hypothesis 57% of the time at $n = 500$ and 82% of the time at $n = 1000$.

2.3.3 Comparison with non-causal approach

We also compared our CTMLE-based test for identifying residues causally related to antibody neutralization to a test developed for random forest-based importance measures. Using the simulation design above, we compared the tests' type I error rates (proportion of tests in which the null hypothesis was rejected for W_{10} and W_{85}) and power under the alternative (proportion of tests in which the null hypothesis was rejected for W_{87}). We found that both tests controlled the type I error rate for the uncorrelated residue (W_{10}); however, for the correlated residue (W_{85}), the type I error for the random forest-based test was inflated relative to 0.05. The inflation remained at larger sample sizes. On the other hand, the random forest-based test exhibited higher power under the alternative (W_{87}).

Sample Size	Methods	W_{10}	W_{85}	W_{87}
500	RF	0.048	0.072	0.746
500	CTMLE	0.029	0.04	0.57
1000	RF	0.056	0.079	0.971
1000	CTMLE	0.041	0.058	0.82

Table 2.3: Proportion of null hypotheses rejected using a random forest (RF)-based test and the proposed CTMLE-based test. W_{10} is a residue that has low correlation with other residues that are causally related to antibody neutralization; W_{85} is a residue that has strong correlation with W_{87}, which is causally related to antibody neutralization.

2.4 Data analyses

2.4.1 CATNAP datasets

The proposed methods were applied to the Compile, Analyze and Tally NAb Panels (CATNAP) database [45]. This database contains information on HIV virus sequences including the HIV envelope AA sequence (the features of interest), other key variables describing structural and biological information (e.g., geographic origin), and a binary measure of whether the virus can be neutralized by a particular antibody (our outcome of interest). We analyzed five antibodies: VRC01, VRC26.08, PGT145, PGT121 and 10-1074. For each antibody, a pre-screening step was applied to exclude AA positions that were nearly constant across all virus sequences. In practice, some residues in Env may exhibit little or no variation. For these residues, we wish to avoid hypothesis testing since there is no hope of garnering sufficient statistical evidence for signal detection. In particular, at each position, AAs that were observed in fewer than 30 total virus sequences were combined into a single AA category. After the combination, only residues still with smallest level count greater than 30 were retained for analysis. Since the number of remaining AA residues that pass this variable screen procedure is still large, we used a Bonferroni correction to control the chance of falsely rejecting null hypotheses. Namely, we tested each individual residue at a significance level of $\alpha/N_{AA}(\text{post-screening})$, where $N_{AA}(\text{post-screening})$ is the number of AA residues passing the screening above.

2.4.2 Results

The number of AA residues that passed the screening process are summarized in Table 2.4 and the plots of $-\log_{10}$(p-value) for each residue are shown in Figure 2.3. For each antibody, the figure highlights functional regions of the gp120 protein that are putatively associated with antibody activity. Our analysis revealed that many significant residues fell within known regions – the V1/V2 loop in particular contained many significant residues – however, there were also many significant residues in gp41, particularly for VRC01.

Antibody	N_{obs}	N_{AA} (pre-screening)	N_{AA} (post-screening)	N_{sig} (%)
VRC01	828	784	328	23 (7.0%)
VRC26.08	407	726	229	3 (1.3%)
PGT145	581	755	294	3 (1.0%)
PGT121	581	755	313	8 (2.6%)
10-1074	581	755	303	8 (2.6%)

Table 2.4: Summary of CATNAP data; N_{obs} = number of virus sequences; N_{AA} = number of amino acid residues

2.4.3 Comparison with other approaches

We compared our results for the VRC01 antibody to the LASSO and random forest approach by feature selection described above. For LASSO-based approach, there were more than 15 multi-level features selected in the CV-selected model, so the inference from the GLM in the held-out data was highly unstable. This illustrates a potential difficulty of this approach, which requires the analyst to make arbitrary, post-hoc modeling decisions to stabilize inference. On the other hand, our proposed CTMLE provides a fully automated procedure that balances type I and II errors.

The random forest-based approach provided greater stability and identified 76 significant residues. However, less than half (16/76) were sites in gp120, which has been identified through other experiments as the likely region of importance for neutralization by VRC01. This result raises the possibility of a high false positive rate of the random forest procedure in this context. Additionally, the associated variable importance measures have

dubious relevance in terms of their scientific interpretation. In contrast, our CTMLE is able to provide biologically meaningful interpretations of estimated parameters.

2.5 Discussion

In this study, we designed a test capable of detecting AA residues along the HIV Env protein that are significantly associated with antibody resistance. To that end, we introduced a CTMLE-based test that is able to cope with the high-dimensional and correlated nature of the data. Our simulation study showed only a relatively modest impact of correlation on performance. In particular, we found the type I error rate was appropriately controlled for null AA residues, regardless of how correlated they were with a true residue. In contrast, the standard TMLE approach tended to have inflated type I error that increased with the strength of correlation. On the other hand, outcome-adaptive approaches tended to better control the type I error of the test. Among outcome-adaptive TMLE estimators, the type I error was generally better controlled when fewer features were included in the GPS model. However, including too few features led to overly conservative tests, where type I error was properly controlled, but power to detect true signals suffered as a result. We found that CTMLE can be used to appropriately determine the best number of features to include resulting in a test that balances type I and type II errors.

An important area for future research is the performance of the method in settings where there is less sparsity in the set of predictors. Understanding the extent to which the CTMLE is able to adapt to this setting is an open question. There is also room for the optimization in this work along other dimensions. First, the method could in theory be extended to other machine learning frameworks beyond random forest that provide suitable rankings of feature importance. Another potential for improving on this work is in considering other multiplicity corrections while performing hypothesis testing, such as Holm's method and Hochberg's method [26, 25]. Finally, the random forest simulation results point to a need

for a more extensive comparison of available methods suited to this problem.

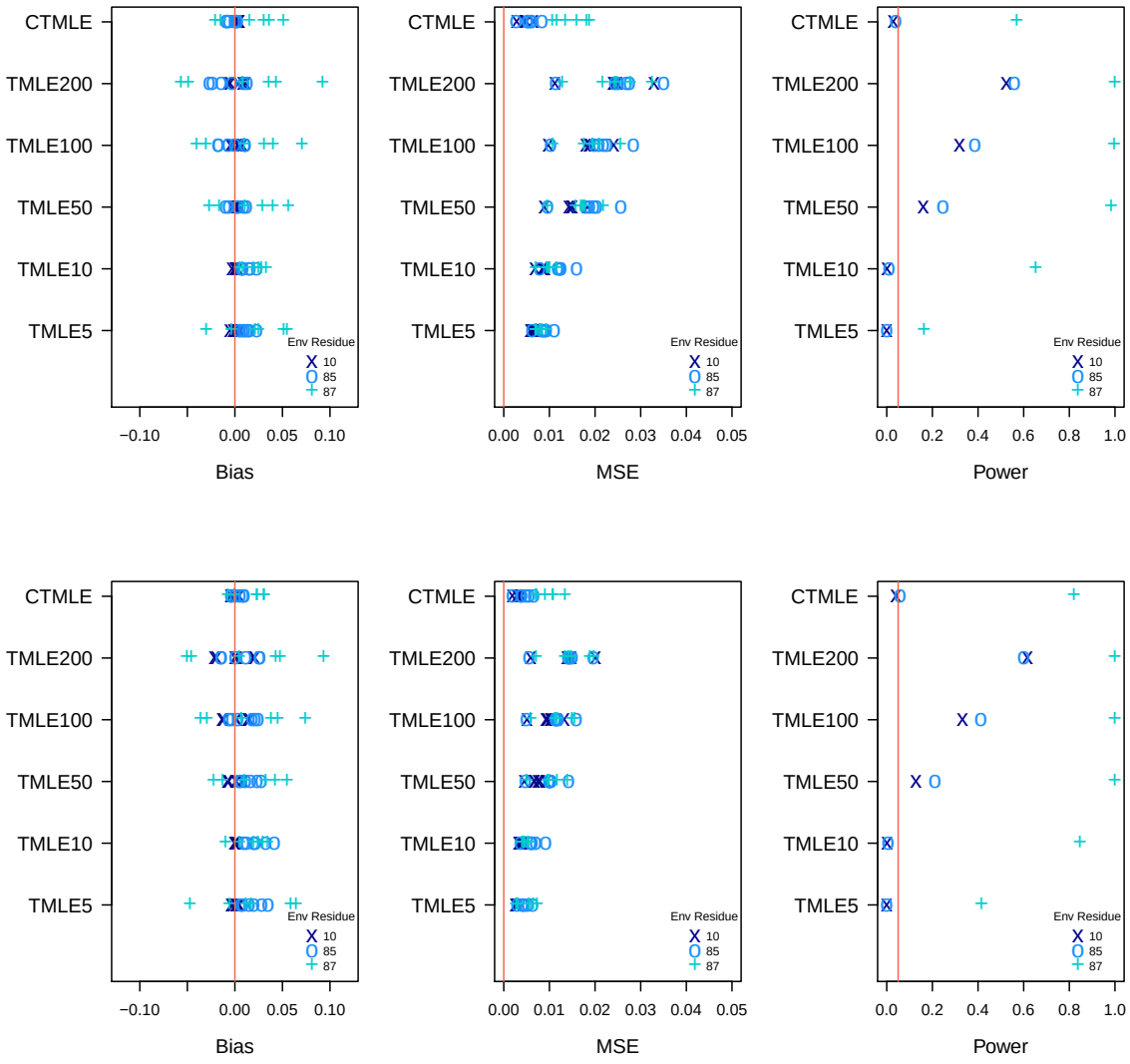

Figure 2.2: **Bias, mean-squared error (MSE), type I error rate for null cases and power for true signals**. Simulation results with sample size of 500 (first row) and 1000 (second row) were visualized for two null cases and one true signal. Comparing with standard TMLE aproach using all 200 features, CTMLE largely reduced bias and MSE, especially for the non-signals; it also maintained a relatively high power with better controlled type I error as good as the best model among five proposed TMLE scenarios. As sample size increases, the power for CTMLE increased from 0.57 to 0.82.

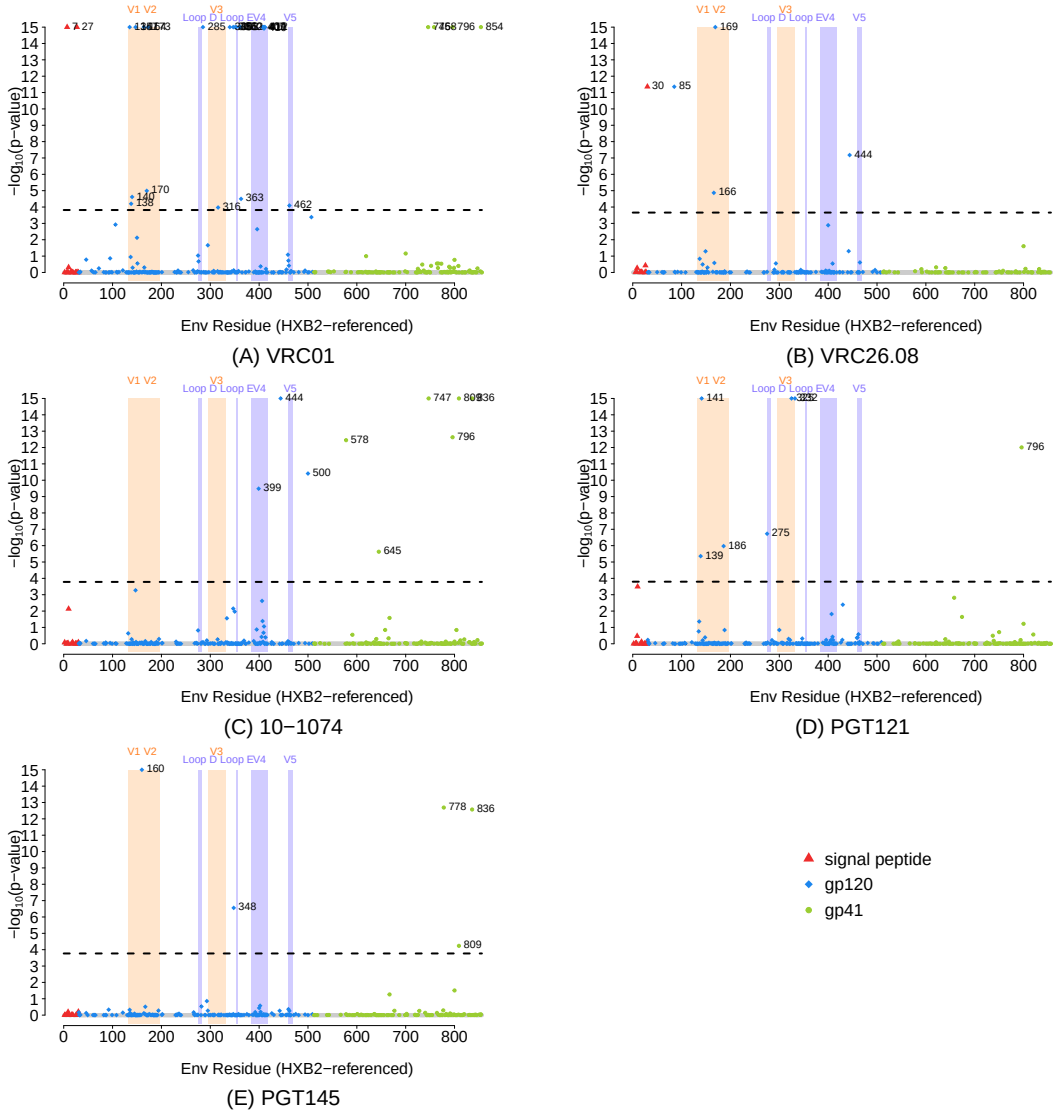

Figure 2.3: **Significant amino acid residues for five antibodies**. Orange highlights denote V1/V2/V3 loop regions that are putatively associated with these antibodies; purple regions highlight additional functional regions of the gp120 protein. Gray points denote residues that did not pass variability screening.

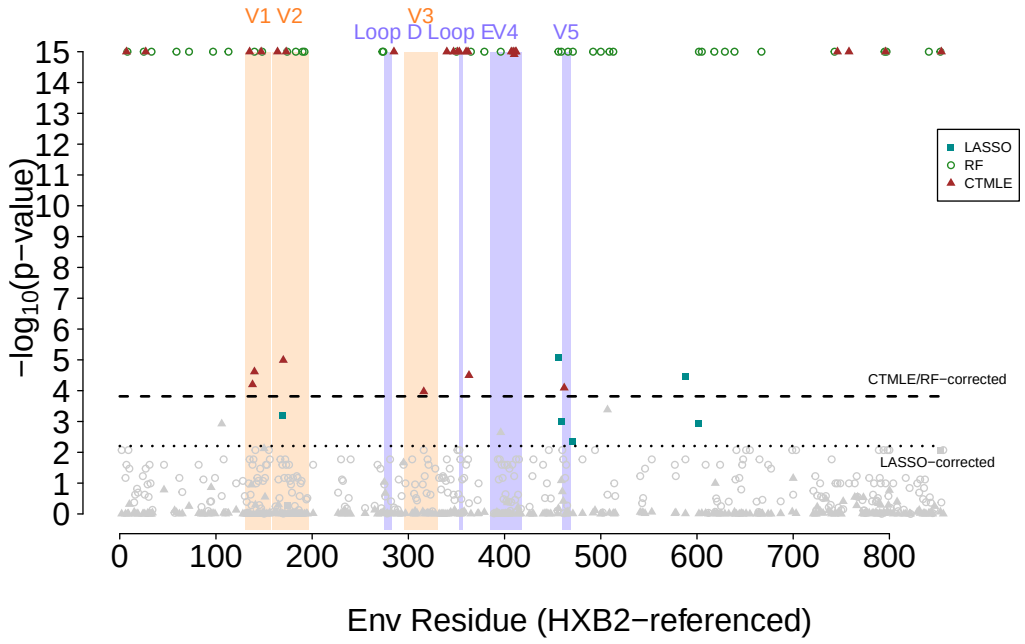

Figure 2.4: **Results for the VRC01 antibody based on three approaches: LASSO, random forests (RF), and CTMLE.** The dashed lines indicate thresholds from a Bonferroni-correction to compensate multiple comparisons.

Chapter 3

Comparing HIV Vaccine Immunogenicity across Trials with Different Populations and Study Designs

3.1 Introduction

In this chapter, we develop a framework that identifies appropriate causal estimands and estimators that can be used to provide standardized comparison of vaccine immunogenicity. We propose estimators of these causal estimands and establish theory that dictates the large sample behavior of the estimators. Our estimators account for two practical difficulties that arise in the study of vaccines. First, we propose methodology that accounts for the fact that the measurement of immune responses across trials may be subject to different sampling designs. For example, HVTN100 and HVTN097 were randomized trials where immune responses were measured in all participants; however, in RV144 and HVTN702 immune responses were measured using case-control sampling. Second, our methodology allows for pooling of trial data to gain efficiency when the same vaccine is evaluated in multiple trials. For example, an identical vaccine was evaluated in HVTN100 and HVTN702 and therefore, we may wish to pool data from these trials when evaluating immunogenicity. We clarify the formal causal assumptions and semiparametric efficiency theory that allows such pooling.

The rest of this paper is organized as follows. In Section 3.2, we detail the estimation and inference procedure using TMLE. Section 3.3 presents extensive simulation studies to evaluate the performance of the proposed estimands in terms of bias, variance, mean squared error (MSE), and the coverage and width of confidence intervals (CIs). In Section 3.4, we apply our method to data from four typical trials in HIV Vaccine Trials Network (HVTN) and compare the proposed estimators to the standard approach.

3.2 Materials and Method

3.2.1 Notation and Data Structure

Vaccine trials can be generally categorized as being phase I/IIa or phase IIb/III design. Phase I/IIa studies are early phase trials, often designed specifically to evaluate immunogenicity of one or several vaccine candidates and/or candidate doses of vaccine. These trials typically have smaller sample sizes, on the order of several hundred participants, and are not powered to provide an assessment of VE on a clinical endpoint of interest. They may include one or several doses of a single vaccine, or one or several variations of a vaccine (e.g., different adjuvants). We use the variable $T \in \mathscr{T} = \{1, 2, \ldots, N_T\}$ to denote an arbitrary numeric label applied to the various trials considered in a particular application. Data in each of these trials contains a possibly categorical variable indicating which of the vaccine formulations/doses a participant receives denoted by the label $A \in \mathscr{A} = \{0, 1, 2, \ldots, N_A\}$. A given vaccine $a \in \mathscr{A}$ could be evaluated in multiple trials; however, in our notation we use only a single, unique label for each vaccine and we denote by $\mathscr{T}_a \subseteq \mathscr{T}$ the trials in which the immunogenicity of vaccine a was evaluated. The observed data also include measurements of one or several immune responses of interest S. In practice S may be a vector, but we focus here only on scalar-valued S, as we can separately apply our methods to each immune response of interest. As a concrete example, we may consider HVTN702, a phase III trial where participants were randomized to receive either an active vaccine or a placebo, and the immune responses of interest included various T cell polyfunctionality scores and IgG binding antibody responses/magnitudes.

Each trial's data will also generally include other participant-level information collected prior to vaccine assignment. The specific baseline characteristics measured may vary across trials, and we introduce $\boldsymbol{W}(t)$ to denote covariates measured in trial $t = \{1, 2, ..., \tau\}$. We use $\boldsymbol{W} = \boldsymbol{W}(1) \cup \boldsymbol{W}(2) \cup ... \cup \boldsymbol{W}(\tau)$ to denote the superset of covariates consisting of all covariates collected in at least one of the trials considered. In our HIV vaccine example,

participants in HVTN100 and 702, participants had their age, gender, body mass index (BMI), region of enrollment and educational level recorded, while in RV144 participants had their age and gender recorded. Thus, in this example $\boldsymbol{W}$ would include five traits that were present in at least one of the three trials.

In addition to vaccine, immune response, and covariates, some trials will also have data available on clinical endpoints of interest. This will almost always be the case for larger phase IIb/III trials that are designed explicitly to evaluate VE. For example, in HVTN702, the primary outcome was time to first detection of HIV-1 infection and this information is recorded for most all participants in the trial. Thus, we can assume that for some trials, the observed data will also include a clinical outcome of interest, which we denote by Y. The outcome could be binary (e.g., indicator of disease by a fixed time-point) or it may be a time-to-event endpoint (e.g., time since vaccination until first occurrence of clinical disease). Our methods apply readily to both situations; however, for simplicity we hence assume Y is binary. In early phase trials, Y may be missing or right-censored for most or all individuals. This missingness pattern has no adverse impact on our developments, since we are primarily interested in comparing S across vaccines and Y occurs after S. If Y is subject to missingness it will be entirely appropriate for our developments to consider this variable as a three-level categorical variable with levels 0, 1, and missing.

An interesting aspect of the design of many vaccine trials is that S may not be measured on every participant. Therefore, we introduce two versions of the data structure that allow us to differentiate between settings where S is measured on every participant in every trial and settings where S is only measured on a subset of participants in at least some of the trials considered. We refer to a datum collected in the former setting as a *full data unit* and in the latter setting as a *observed data unit*. Explicitly considering the full data unit in this setting is useful mathematically for describing requisite assumptions for identification of our causal estimands of interest. In the full data setting, for each participant in each trial, we record $X = (T, A, \boldsymbol{W}, S, Y) \sim P_X$, while recalling that without loss of generality, Y will

generally be coded as right-censored or missing for most or all individuals enrolled in early phase trials. In our notation, we use P_X to denote the probability distribution X, which is assumed to fall in a statistical model $\mathscr{M}_X$ that is nonparametric up to certain assumptions detailed in Section 3.2.3 below. We use E_X to denote expectation of a random variable under sampling from P_X.

We now turn to the *observed data unit*, where S may not be measured on some individuals by design due to time and budgetary considerations. Many phase IIb/III trials employ a two-stage sampling design to determine the subset of participants in which S should be measured. In HVTN702, a case-control design was used, wherein vaccinated participants who were observed to be HIV-1 infected during follow up uniformly had S measured. A matched set of controls also had S measured. Thus, while 1168 participants were enrolled in HVTN702, S was only measured on 130 of these individuals. To accommodate the potential for the presence of two-stage sampling, we introduce the *observed data unit* $O = (T, A, \boldsymbol{W}, \Delta, \Delta S, Y) \sim P$, which is a *coarsened* version of the full data unit X. A typical observed data unit includes T, A, $\boldsymbol{W}$ and Y (possibly subject to missingness) as above; however, the immune response S is measured only in a subset of participants. The random variable Δ takes value 1 if the immune response S is measured and zero otherwise. In the data unit O, without loss of generality we represent the observed value of S by ΔS, thereby arbitrarily recording a value of 0 for S in individuals not selected for two-phase sampling. We note that for early phase trials generally we will have $\Delta = 1$ for all participants, while for late phase trials, $\Delta = 1$ for only a subset of participants. The statistical model $\mathscr{M}$ for P is implied by the model for the distribution of the full data unit P_X and the model for the sampling variable Δ, where these sampling probabilities are generally known by design. We use E to denote the expectation of random variable under sampling for P.

Table 3.1 provides an example visualization of the type of coarsened data that is used in our motivating example. In this example, our data set consists of data from four trials pooled into a single data set. Our two covariates $\boldsymbol{W}$ of interest are categorical participant

Table 3.1: Example data set showing data structure for typical HIV vaccine trials.

T	Y	S	Δ	A	W_1	W_2
HVTN702	1	1.2	1	1	≤ 20	F
HVTN702	0	-	0	1	≤ 20	F
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
HVTN702	0	0	1	0	> 20	M
HVTN702	0	-	0	0	> 20	M
RV144	1	-	0	0	≤ 20	F
RV144	0	2.1	1	2	> 20	M
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
RV144	0	0	1	0	> 20	M
RV144	0	-	0	2	≤ 20	F
HVTN100	-	1.8	1	3	≤ 20	F
HVTN100	-	2.8	1	3	≤ 20	F
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
HVTN100	-	4.2	1	3	> 20	M
HVTN100	-	2.8	1	3	≤ 20	F
HVTN097	-	0	1	1	≤ 20	F
HVTN097	-	0	1	1	≤ 20	F
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
HVTN097	-	0	1	1	≤ 20	F
HVTN097	-	1.3	1	1	> 20	M

age (W_1) and sex (W_2). There are two late phase trials, HVTN702 and RV144 and two early phase trials HVTN100 and HVTN097. In the late phase trials, HIV-1 infection status Y is recorded for all participants (for simplicity, we present an idealization of the actual data where we ignore right-censoring of Y). However, the immune response S of interest is only measured in a subset of participants in these trials, as indicated by rows where $\Delta = 1$; S is missing for all rows in which $\Delta = 0$. On the other hand, in the early phase trials, S is measured for everyone, while Y is generally missing. At times, we will refer back to Table 3.1 to make concrete our general estimation strategies.

3.2.2 Causal Estimands

Traditionally, the average immune response induced by each vaccine is estimated in each trial separately. This approach targets the estimand $\mu_a := E_X(S \mid A = a, T = t)$ for various combinations of a and t. However, in some situations, there may be certain components of $\boldsymbol{W}$ that are correlated with both trial enrollment T, as well as immune responses S. For example, age distributions of participants may vary across trials, while age also commonly correlates with the magnitude of vaccine-induced immune responses. Thus, a comparison of μ_a and $\mu_{a'}$ for two vaccine candidates a and a' evaluated in different trials t and t' may be biased for a causal estimand that is truly of interest.

To address this concern, we propose a causal framework to provide such comparisons in an appropriate way. In particular, we can consider a counterfactual variable $S(a)$ that corresponds to the immune response that would be observed if an individual were given vaccine a. We assume that causal consistency holds and that there is no interference between individuals [37]. Both assumptions are reasonable in the present context, where causal consistency stipulates that there are not "multiple formulations" of a single vaccine. This assumption is generally reasonable for most vaccines, where often a key goal of preclinical vaccine development process is developing consistent manufacturing processes to ensure comparable vaccines across lots. No interference is also likely to be plausible in the present context as the immune response of one individual is unlikely to depend on vaccines received by other individuals in the study.

In this counterfactual scenario, it is possible for all individuals who could potentially enroll in any of the trials to receive any of the N_A vaccines considered. Thus, we can conceptualize a counterfactual data unit $\mathbb{X} = (T, \boldsymbol{W}, \{S(a), Y(a) : a \in \{1, \ldots, N_A\}\}) \sim P_{\mathbb{X}}$, where for completeness we define $Y(a)$ as the counterfactual clinical endpoint that would be observed under vaccination with a, though this quantity does not play a role in our development. As above, we denote by $E_{\mathbb{X}}$ expectation of a random variable under $P_{\mathbb{X}}$.

We are ultimately interested in comparing immunogenicity, for example, by comparing

the average value of $S(a)$ vs. $S(a')$ for vaccines a, a' that were evaluated across different trials. However, when these vaccines are evaluated across different trials that enroll from different populations, there are several such comparisons that could be of interest. In the context of HIV vaccines, a series of trials were conducted in several countries across several years. As described in the introduction, in our motivating example, the population of HVTN702 was of primary interest, as our goal is to compare the immunogenicity across vaccines to aid in the interpretation of the null signal in the primary vaccine efficacy analysis of the HVTN702. Thus, we may be interested in understanding whether and how the immunogenicity of the vaccine formulation studied in the earlier RV144 trial compares to the formulation studied in HVTN702, while making this comparison *in the HVTN702 trial population*. That is, we are asking a hypothetical question about the immunogenicity that *would have been observed* had we evaluated the RV144 vaccine alongside the HVTN702 vaccine, *in the HVTN702 study population*. Using the labels from Table 3.1, this estimand would be denoted $E_{\mathbb{X}}[S(1) - S(2) \mid T = \text{HVTN702}]$. We label this type of causal estimand a *standardized* comparison of immunogenicity.

While our motivating example focuses on a setting where a single trial's population is of interest, more generally we could consider standardized comparisons of the form $E_{\mathbb{X}}[S(a) - S(a') \mid T \in \mathscr{T}_{\text{ref}}]$, where $\mathscr{T}_{\text{ref}} \subseteq \mathscr{T}$ may include multiple trials. We refer to $\mathscr{T}_{\text{ref}}$ as the *referent trial(s)* to which we are standardizing our comparison. The choice of referent trial should be dictated by the scientific context. While we generally expect that $\mathscr{T}_{\text{ref}}$ will consist of a single trial, in some situations we may wish to include multiple trials in our referent. For example, if vaccines a and a' are evaluated in trials that enroll from very similar or identical populations, then we may wish for $\mathscr{T}_{\text{ref}} = \mathscr{T}_a \cup \mathscr{T}_{a'}$. A trivial situation where this might occur is when vaccines a and a' are evaluated in the same trial and we are interested in inference on their immunogenicity in that trial's study population. However, we may also have settings where vaccines are evaluated in different trials, but the distribution of common baseline covariates are largely similar among trials $\mathscr{T}_a$ and $\mathscr{T}_{a'}$. This

could happen when vaccines are evaluated at the same study sites using trials with similar enrollment criteria. For example, multiple COVID-19 vaccines were evaluated in 2020 using randomized clinical trials that leveraged shared study sites as part of the COVID-19 Prevention Trials Network [13]. These trials enrolled from largely similar populations and we may wish to compare immunogenicity of the various vaccines across the pooled study population. In the absence of effect heterogeneity by covariates, inference on this quantity may enjoy greater precision than inference based on an estimand standardized to either $\mathscr{T}_a$ or $\mathscr{T}_{a'}$ alone.

Remark: Another potential setting is one where there exists a common referent population that is not sampled directly from any of the observed trials. For example, we may wish to compare immunogenicity of two vaccines in an age- and sex-standardized way against a known referent population distribution (e.g., the age-, sex-distribution of all individuals in a particular country or available as part of the electronic medical records for some healthcare system). Letting $Q_{\boldsymbol{W}^*}$ denote the cumulative distribution function of $\boldsymbol{W}$ in the referent population, we may be interested in $\int E_{\mathbb{X}}[S(a') - S(a) \mid \boldsymbol{W} = \boldsymbol{w}] dQ_{\boldsymbol{W}^*}(\boldsymbol{w})$. Such a comparison may be particularly useful for national modeling studies of vaccination impact, where we need to understand immunogenicity at the level of a specific population that is not specific to one trial.

3.2.3 Identification of standardized immunogenicity using full data

To identify the standardized immunogenicity comparison described in our motivating example, we require certain *causal* assumptions regarding the distribution of $\mathbb{X}$, in addition to assumptions pertaining to the sampling design of S as encoded in the distribution of O. Key to both sets of assumptions is the consideration of which baseline covariates are available across the various trials.

We introduce the general notation $\boldsymbol{W}_{\cap}(\mathscr{T}_0)$ to denote covariates that are available across *all* of a given set of trials $\mathscr{T}_0 \subseteq \mathscr{T}$. Thus, $\boldsymbol{W}_{\cap}(\mathscr{T}_{\text{ref}}) \subseteq \boldsymbol{W}$ refers to the covariates common

to all referent trial(s) and $\boldsymbol{W}_{\cap}(\mathscr{T}_a)$ denotes covariates common to all trials where vaccine a is evaluated. We denote by $\boldsymbol{W}_{\cap}(\mathscr{T}_{\text{ref}}) \cap \boldsymbol{W}_{\cap}(\mathscr{T}_a)$ the set of covariates that are available in all referent trials and all trials where vaccine a is evaluated. This set of covariates is particularly important for identification. As we presently show, we must be able to identify a subset of these covariates that is sufficient to control for differences in counterfactual immunogenicity between the individuals receiving vaccine a and individuals in the referent population. We make the simplifying assumption that such covariates *must* be available in all of the trials where the immunogenicity of vaccine a is actually measured, so that we can identify the vaccine's expected immunogenicity conditional on this set of covariates. Moreover, we also need the *same set of covariates* to be available in the referent trial(s) so that the covariate-conditional immunogenicity can be properly standardized to the referent trial. Future work will be devoted to identifying under the weaker assumption that covariates are only available in at least one trial where vaccine a is evaluated.

Formally, identification of $\psi_{\mathbb{X}}(a) = E_{\mathbb{X}}[S(a) \mid T \in \mathscr{T}_{\text{ref}}]$ for an arbitrary vaccine a using the full data requires the following assumptions.

(A1) *Ignorability of trial enrollment and vaccine assignment conditional on common covariates*. There exists a set of common baseline covariates $\boldsymbol{W}_S \subseteq \boldsymbol{W}_{\cap}(\mathscr{T}_{\text{ref}}) \cap \boldsymbol{W}_{\cap}(\mathscr{T}_a)$ such that:

(A1.1) $S(a) \perp A \mid \boldsymbol{W}_S$

(A1.2) $S(a) \perp T \mid \boldsymbol{W}_S$

(A2) *Positivity of vaccine assignment*

(A2.1) $P_X\{P_X(A = a \mid \boldsymbol{W}_S) > 0 \mid T \in \mathscr{T}_{\text{ref}}\} = 1$

Assumption (A1.1) stipulates that we must be able to identify a set of covariates that are measured in both the referent trial and the trial(s) where vaccine a is evaluated such that conditional on this set of covariates, the vaccine which a participant is observed to

receive provides no additional information about their potential outcome $S(a)$. This condition will generally hold by design if vaccines are randomly assigned in all trials included as part of $\mathscr{T}_a$. However, if $\mathscr{T}_a$ includes one or more observational studies, this assumption would require additional scrutiny. Assumption (A1.2) stipulates that conditional on $\boldsymbol{W}_S$, the particular trial in which vaccine a was evaluated provides no additional information about the potential outcome $S(a)$. Generally, we can think about two sub-assumptions that are needed to satisfy this assumption. First, there can not be a direct effect of trial on vaccine immunogenicity. This assumption would be violated if, for example, one trial in $\mathscr{T}_a$ had inappropriate cold storage procedures for a vaccine thereby causing weakened immunogenicity of the vaccine. Second, we require that $\boldsymbol{W}_S$ includes all characteristics that are related to both vaccine immunogenicity and that may differ across trial populations. For example, consider a scenario where certain compositions of the gut microbiome have a positive impact on vaccine immunogenicity and microbiome data are not available as part of $\boldsymbol{W}_S$. If microbiome composition differs across trials in $\mathscr{T}_a$, then assumption (A1.2) would be violated. Graphical approaches may be useful for scrutinizing this assumption in each specific scientific context. We remark that our notation $\boldsymbol{W}_S$ indicates that the choice of covariates may differ depending on the particular immune response that is being studied, as different responses may have different biological drivers. The choice of covariates $\boldsymbol{W}_S$ may also differ depending on the particular vaccine a and the particular choice of referent trial $\mathscr{T}_{\text{ref}}$. However, for simplicity we have elected to suppress this dependency in our notation.

Assumption (A2.1) stipulates that there is a positive probability of receiving vaccine a for all values of $\boldsymbol{W}_S$ that are observable in $\mathscr{T}_{\text{ref}}$. This assumption would be violated if, for example, there were certain combinations of covariates that are observable in the referent trials $\mathscr{T}_{\text{ref}}$, but not in any of the trials in which vaccine a was studied. This condition could be scrutinized empirically using standard methods for evaluating propensity score overlap, for example by evaluating an estimate of $P_X(A = a \mid \boldsymbol{W}_S)$ using observations in $\mathscr{T}_{\text{ref}}$ [2].

Theorem 1. *If (A1) and (A2) hold, then*

$$\psi_{\mathbb{X}}(a) = E_X\left[E_X(S \mid A = a, \boldsymbol{W}_S) \mid T \in \mathscr{T}_{ref}\right] .$$

A detailed proof can be found in Appendix A. We hence use $\psi_X(a) = E_X[E_X(S \mid A = a, \boldsymbol{W}_S) \mid T \in \mathscr{T}_{\text{ref}}]$ to refer to the identifying estimand as distinct from the causal estimand $\psi_{\mathbb{X}}(a)$. The implication of Theorem 1 is that if (A1) and (A2) hold then $\psi_{\mathbb{X}}(a) = \psi_X(a)$ and a causal standardized immunogenicity comparison is possible using data sampled from P_X. However, even in the ideal context where $\mathscr{T}_a$ consists of only randomized trials, assumption (A1.2) may yet be considered unreasonable. For example, the various trials in $\mathscr{T}_a$ may collect different sets of key covariates, rendering this assumption difficult to satisfy based on the set of covariates common to $\mathscr{T}_{\text{ref}}$ and $\mathscr{T}_a$. In this case, $\psi_X(a)$ does not have a causal interpretation. Nevertheless, we argue that a comparison of $\psi_X(a)$ and $\psi_X(a')$ still retains a useful non-causal interpretation as a covariate-adjusted comparison of vaccines a and a', standardizing the set of available common covariates to their distribution in the referent trial(s). So long as $\boldsymbol{W}_S$ contains at least some covariates that are prognostic of immune response and whose distributions differ across trials, we argue that a comparison of $\psi_X(a)$ and $\psi_X(a')$ may still be preferred over a naïve estimand that compares μ_a and $\mu_{a'}$.

3.2.4 Identification of standardized immunogenicity using observed data

We now describe how $\psi_X(a)$ can be identified in the coarsened data setting, where we are sampling data from P rather than P_X. In a particular trial t, sampling probabilities for S could be determined based on A (e.g., we may over-sample vaccine recipients and under-sample placebo recipients), Y (e.g., it is common to sample all cases and only a subset of the remaining trial participants), and/or a subset of available covariates $\boldsymbol{W}(t)$ (e.g., we may over-sample minority or elderly populations to ensure appropriate representation in

the observed data). We denote by $\boldsymbol{W}_\Delta(t) \subseteq \boldsymbol{W}(t)$ the set of covariates, if any, that are used to determine sampling probabilities in trial t. Here, to simplify the exposition, we make the simplifying assumption that all trials in $\mathscr{T}_a$ use two-stage sampling and that the covariates used for sampling, $\boldsymbol{W}_\Delta(t)$ are the same for all such trials. We refer to this set of covariates as $\boldsymbol{W}_\Delta$, suppressing the dependence on a for simplicity. In future work, we will demonstrate how this assumption may be relaxed to allow different sampling designs across $\mathscr{T}_a$; we expect this generalization to be straightforward.

To identify $\psi_X(a)$ using the observed data we require the following assumptions.

(A3) *Coarsening at random*

(A3.1) $S \perp \Delta \mid T, A, \boldsymbol{W}_\Delta, Y$

(A4) *Positivity of sampling probability*. For all $t \in \mathscr{T}_a$,

(A4.1) $P\{P(\Delta = 1 \mid T = t, A = a, \boldsymbol{W}_\Delta, Y) > 0 \mid T \in \mathscr{T}_{\text{ref}}\} = 1$

Assumption (A3.1) stipulates that given $(T, A, \boldsymbol{W}_\Delta, Y)$ the probability of having immune responses measured cannot depend on the underlying immune response itself. Sampling probabilities are generally selected a-priori by design in late phase vaccine trials, so we expect this assumption will typically be satisfied. If instead, trials are designed such that immune responses are measured subject to some form of convenience sampling (e.g., participants can self-select into an immunogenicity sub-study), then this assumption would require further scrutiny. Assumption (A4.1) stipulates that there is a positive probability of sampling immune responses for measurement for each available profile existing in $\mathscr{T}_{\text{ref}}$, which again can generally be ensured by design. We define $\boldsymbol{W}_{\Delta,S} = \boldsymbol{W}_S \cup \boldsymbol{W}_\Delta$ to be the union of covariates required to satisfy (A1) and (A3). We have the following identification result for $\psi_X(a)$.

Theorem 2. *If Assumptions (A3)-(A4) hold then*

$$\psi_X(a) = E\{E[E(S \mid \Delta = 1, A = a, T \in \mathscr{T}_a, Y, \boldsymbol{W}_{\Delta,S}) \mid A = a, T \in \mathscr{T}_a, \boldsymbol{W}_S] \mid T \in \mathscr{T}_{ref}\} .$$

3.2.5 Towards estimation: efficiency theory for identifying estimands

In this section, we provide the efficient influence function of $\psi_X(a)$ and $\psi(a)$ in models that assume (A1)-(A4). We recall that an estimator's *influence function* is a function of the data unit that has mean zero and finite variance. In particular, an estimator $\psi_n(a)$ of $\psi(a)$ is said to have influence function D if $\psi_n(a) = \psi(a) + n^{-1}\sum_{i=1}^{n} D(O_i) + o_P(n^{-1/2})$. Influence functions are particularly useful for so-called *regular* estimators, as they can also be used to characterize the efficiency bound of all such estimators of a given parameter. The influence function of the regular estimator with the smallest asymptotic variance is called the efficient influence function. Influence functions are often indexed by so-called *nuisance parameters*, parameters of the data generating distribution that are not directly of interest, but are useful for constructing and studying the large sample behavior of estimators of the estimand of interest. Thus, influence functions can provide hints as to what quantities must be estimated to generate estimates with desirable large sample behavior.

We introduce some additional notation to represent the nuisance parameters indexing our efficient influence function. We define $\bar{Q}_X(\boldsymbol{W}_{i,S}) = E_X(S \mid A = a, \boldsymbol{W}_S = \boldsymbol{W}_{i,S})$ as the conditional mean immune response, $g_A(a \mid \boldsymbol{W}_{i,S}) = P_X(A = a \mid \boldsymbol{W}_S = \boldsymbol{W}_{i,S})$ as the conditional probability of vaccine a given covariates $\boldsymbol{W}_{i,S}$, $g_T(\mathscr{T}_0 \mid \boldsymbol{W}_{i,S}) = P_X(T \in \mathscr{T}_0 \mid \boldsymbol{W}_S = \boldsymbol{W}_{i,S})$ as the conditional probability of enrollment in one of the trials in a given set $\mathscr{T}_0 \subseteq \mathscr{T}$ given covariates $\boldsymbol{W}_{i,S}$, and $g_T(\mathscr{T}_{\text{ref}}) = P_X(T \in \mathscr{T}_{\text{ref}})$ as the marginal probability of enrollment in one of the trials in $\mathscr{T}_{\text{ref}}$.

Theorem 3. *The efficient influence function for $\psi_X(a)$ in a model for P_X that only assumes (A1)-(A2) is*

$$D_X(X_i) = \frac{\mathbb{1}_a(A_i)}{g_A(a \mid \boldsymbol{W}_{i,S})} \frac{g_T(\mathscr{T}_{ref} \mid \boldsymbol{W}_{i,S})}{g_T(\mathscr{T}_{ref})} \{S_i - \bar{Q}_X(\boldsymbol{W}_{i,S})\} + \frac{\mathbb{1}_{\mathscr{T}_{ref}}(T_i)}{g_T(\mathscr{T}_{ref})} \{\bar{Q}_X(\boldsymbol{W}_{i,S}) - \psi_X(a)\} .$$

We can also define the efficient influence function for $\psi(a)$, which is indexed by the

following additional nuisance parameters:

$$\bar{Q}_2(Y_i, \boldsymbol{W}_{i,\Delta,S}) = E(S \mid \Delta = 1, A = a, T \in \mathscr{T}_a, Y = Y_i, \boldsymbol{W}_{\Delta,S} = \boldsymbol{W}_{i,\Delta,S}) \ ,$$

$$\bar{Q}_1(\boldsymbol{W}_{i,S}) = E[\bar{Q}_2(Y, \boldsymbol{W}_{\Delta,S}) \mid A = a, T \in \mathscr{T}_a, \boldsymbol{W}_S = \boldsymbol{W}_{i,S}] \ ,$$

$$g_\Delta(1 \mid T_i, A_i, Y_i, \boldsymbol{W}_{i,\Delta,S}) = P(\Delta = 1 \mid T = T_i, A = A_i, Y = Y_i, \boldsymbol{W}_{\Delta,S} = \boldsymbol{W}_{i,\Delta,S}) \ .$$

Theorem 4. *The efficient influence function for $\psi(a)$ in a model for P that assumes (A1)-(A4) is*

$$\begin{aligned} D(O_i) = &\frac{\mathbb{1}_1(\Delta)}{g_\Delta(1 \mid T_i, A_i, Y_i, \boldsymbol{W}_{i,\Delta,S})} \frac{\mathbb{1}_a(A_i)}{g_A(a \mid \boldsymbol{W}_{i,S})} \frac{g_T(\mathscr{T}_{ref} \mid \boldsymbol{W}_{i,S})}{g_T(\mathscr{T}_{ref})} \{S_i - \bar{Q}_2(Y_i, \boldsymbol{W}_{i,\Delta,S})\} \\ &+ \frac{\mathbb{1}_a(A_i)}{g_A(a \mid \boldsymbol{W}_{i,S})} \frac{g_T(\mathscr{T}_{ref} \mid \boldsymbol{W}_{i,S})}{g_T(\mathscr{T}_{ref})} \{\bar{Q}_2(Y_i, \boldsymbol{W}_{i,\Delta,S}) - \bar{Q}_1(\boldsymbol{W}_{i,S})\} \\ &+ \frac{\mathbb{1}_{\mathscr{T}_{ref}}(T_i)}{g_T(\mathscr{T}_{ref})} \{\bar{Q}_1(\boldsymbol{W}_{i,S}) - \psi(a)\} \ . \end{aligned}$$

3.2.6 Targeted minimum loss estimation

The form of the efficient influence function suggests a natural targeted minimum loss-based estimation (TMLE) approach involving sequential regression [35, 3, 42, 41]. TMLE in general consists of two major steps that are sometimes implemented iteratively. In the first step, estimators of *nuisance parameters* indexing the efficient influence function are obtained. TMLE is agnostic as to how such parameters are estimated, though regression stacking or super learning is commonly used towards this end in practice [43]. The second step of TMLE improves by using empirical risk minimization in a low-dimensional parametric model to simultaneously (i) improve the fit of initial nuisance parameter estimates and (ii) ensure that, at the end of the TMLE procedure, the so-called *efficient influence function estimating equation* is solved. For example, if we denote by D_n the efficient influence function presented in Theorem 2 but where nuisance parameters are replaced by estimated counterparts, then the second step of TMLE ensures that $n^{-1}\sum_{i=1}^{n} D_n(O_i) \approx 0$.

In many problems, including the present, the second step of TMLE can be performed using a simple univariate logistic regression and maximum likelihood estimation, as described below.

A TMLE for $\psi(a)$ may be implemented in the following specific steps.

1. **Estimate the probability of enrollment in referent trial(s) given covariates.** To estimate $g_T(\mathscr{T}_{\text{ref}} \mid \boldsymbol{W}_S)$, we can use data from all observed trials to fit a regression with the outcome equal to $\mathbb{1}_{\mathscr{T}_{\text{ref}}}(T)$ and predictors $\boldsymbol{W}_S$. This regression could be estimated using approach that is appropriate for binary outcome regression. Denote the estimate by $g_{n,T}(\mathscr{T}_{\text{ref}} \mid \cdot)$ and define the estimated marginal probability of enrollment in $\mathscr{T}_{\text{ref}}$ as $g_{n,T}(\mathscr{T}_{\text{ref}}) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\mathscr{T}_{\text{ref}}}(T_i)$.

2. **Estimate the pooled probability of receiving vaccine a given covariates.** To estimate $g_A(a \mid \boldsymbol{W}_S)$, we can use all the data to fit a regression with the outcome equal to $\mathbb{1}_a(A)$, an indicator of receiving vaccine a (versus any other vaccine). The predictors of the regression are $\boldsymbol{W}_S$. This regression could be estimated using approach that is appropriate for binary outcome regression. Denote the estimate by $g_{n,A}$.

3. **Compute sampling probabilities for each individual.** Next, we need to evaluate sampling probabilities $g_\Delta(1 \mid T_i, A_i, Y_i, \boldsymbol{W}_{i,\Delta,S})$ for each individual who received vaccine a and who were enrolled in one of the trials included in $\mathscr{T}_a$. These probabilities are generally known by design; if unknown, then they could be estimated separately for each trial in $\mathscr{T}_a$ using regression of the binary outcome Δ on predictors $A, Y, \boldsymbol{W}_{\Delta,S}$. Denote by $g_{n,\Delta}$ the estimates (or true values) of the conditional sampling probabilities.

4. **Estimate vaccine-specific conditional mean immunogenicity given sampling and other covariates.** To obtain an estimate of $\bar{Q}_2$, we can use data from individuals who received vaccine a across all trials in $\mathscr{T}_a$ to fit a regression with the outcome S and

predictors $Y, \boldsymbol{W}_{\Delta,S}$. As above, any suitable regression technique can be used and we denote by $\bar{Q}_{n,2}$ the estimate of the conditional mean immunogenicity.

5. **Target the vaccine-specific conditional mean immunogenicity given sampling and other covariates.** For simplicity, suppose $S \in (0,1)$. If this assumption does not hold, then S can be re-scaled to fall in this interval and the same approach can be adopted [21]. Using all individuals with measured immune response $\Delta = 1$ that received vaccine a in trials $\mathscr{T}_a$, fit a logistic regression with outcome S, an offset equal to $\text{logit}[\bar{Q}_{n,2}(Y, \boldsymbol{W}_{\Delta,S})]$, and a single covariate, defined as

$$H_{n,2}(T,A,Y,\boldsymbol{W}_{\Delta,S}) = \frac{g_{n,T}(\mathscr{T}_{\text{ref}} \mid \boldsymbol{W}_S)}{g_{n,\Delta}(1 \mid T,A,Y,\boldsymbol{W}_{\Delta,S}) g_{n,A}(a \mid \boldsymbol{W}_S) g_{n,T}(\mathscr{T}_{\text{ref}})} .$$

Note that this regression model for $\bar{Q}_2$ has a single coefficient β_2 and the model can be expressed as $\bar{Q}_{2,\beta_2} = \text{expit}[\text{logit}(\bar{Q}_{n,2}) + \beta_2 H_{n,2}], \beta_2 \in \mathbb{R}$. Let $\beta_{n,2}$ denote the maximum likelihood estimate of β_2 and define $\bar{Q}^*_{n,2}$ as an estimate of $\bar{Q}_{2,\beta_2}$.

6. **Estimate vaccine-specific conditional mean immunogenicity excluding sampling covariates.** To estimate $\bar{Q}_1$, we regress the pseudo-outcome $\bar{Q}^*_{n,2}(T,A,Y,\boldsymbol{W}_{\Delta,S})$ onto $\boldsymbol{W}_S$ using observations that received vaccine a. As above, any suitable regression technique can be used and we denote by $\bar{Q}_{n,1}$ the estimate of conditional mean immunogenicity, now conditioning *only* on baseline covariates $\boldsymbol{W}_S$.

7. **Target the vaccine-specific conditional mean immunogenicity excluding sampling covariates.** For simplicity, we again suppose that our initial estimates obtained in the previous step are such that $\bar{Q}_{n,1}(\boldsymbol{W}_S) \in (0,1)$ for all $\boldsymbol{W}_S$, while re-scaling can again be applied as needed. Now, using *all* individuals that received vaccine a, fit a logistic regression with outcome $\bar{Q}^*_{n,2}(T,A,Y,\boldsymbol{W}_{\Delta,S})$, an offset equal

to logit$[\bar{Q}_{n,1}(\boldsymbol{W}_S)]$, and a single covariate, defined as

$$H_{n,1}(\boldsymbol{W}_S) = \frac{g_{n,T}(\mathscr{T}_{\text{ref}} \mid \boldsymbol{W}_S)}{g_{n,A}(a \mid \boldsymbol{W}_{i,S}) g_{n,T}(\mathscr{T}_{\text{ref}})} \ .$$

Note that this regression model for $\bar{Q}_1$ has a single coefficient β_1 and the model can be expressed as $\bar{Q}_{1,\beta_1} = \text{expit}[\text{logit}(\bar{Q}_{n,1}) + \beta_1 H_{n,1}], \beta_1 \in \mathbb{R}$. Let $\beta_{n,1}$ denote the maximum likelihood estimate of β_1 and define $\bar{Q}^*_{n,1}$ as an estimate of $\bar{Q}_{1,\beta_1}$.

8. **Construct the final TMLE estimate**. The final estimate is

$$\psi^*_n(a) = \frac{1}{\sum_j \mathbb{1}_{\mathscr{T}_{\text{ref}}}(T_j)} \sum_{i=1}^{n} \mathbb{1}_{\mathscr{T}_{\text{ref}}}(T_i) \bar{Q}^*_{n,1}(\boldsymbol{W}_{i,S}) \ .$$

3.2.7 Hypothesis Testings and Confidence Intervals

Under standard regularity conditions [23], our TMLE estimator $\psi_n(a)$ is asymptotically linear and locally efficient. Additionally, the central limit theorem indicates that this estimator converges to a random variable with a mean-zero Gaussian distribution and variance equal to the variance of $D(O)$. This limiting distribution immediately suggests a closed-form, Wald-type confidence intervals (CIs) and hypothesis tests. Specifically, we can construct an asymptotically valid $(1-\alpha)$ Wald-type confidence interval as $\psi_n(a) \pm z_{(1-\alpha/2)} \hat{\sigma}_n / \sqrt{n}$, where $\hat{\sigma}^2_n$ is the empirical variance of $D(O)$ evaluated at the estimated nuisance parameters and $z_{(1-\alpha/2)}$ is the $1-\alpha/2$ quantile of a standard normal distribution. These asymptotic results can also be extended to construct Wald-style hypothesis test for differences between immunogenicity of different vaccines by using the delta method and carefully defined contrast matrices.

3.3 Simulation Studies

We evaluated the proposed estimators via two simulation studies. In the first, we design a scenario that is common observed in vaccine trials where we wish to compare immunogenicity of vaccines across early and late phase trials. In the second, we tailor the data generating mechanism specifically to the HIV vaccine setting described above. In each simulation study, we evaluate estimators in terms of bias, variance, mean squared error (MSE), coverage probability of 95% Wald-type confidence intervals and mean width of confidence intervals.

3.3.1 Comparing within and across early and late phase trials

In this simulation, we wished to compare the immunogenicity of vaccines evaluated in two different studies where the studies had an imbalance of key covariates. In this context, we considered three scenarios: (i) comparing vaccines evaluated in separate early phase trials; (ii) comparing vaccines where one is evaluated in an early phase trial and the other in a late-phase trial that used two-phase sampling for measurement of immune responses; (iii) comparing vaccines evaluated in separate late phase trials that both used two-phase sampling for measurement of immune responses. In each setting, we simulated two binary covariates $W_1 \mid T \sim Bernoulli(0.65+0.15\mathbb{1}_2(T))$ and $W_2 \mid T \sim Bernoulli(0.5-0.2\mathbb{1}_2(T))$. We use $A=1$ to denote the vaccine evaluated in trial $T=1$ and $A=2$ to denote the vaccine evaluated in trial $T=2$. Both trials 1 and 2 were simulated to have 1:1 randomization to either their respective active vaccines or a control vaccine (arbitrarily labeled $A=3$). The immune response was simulated as $S \mid A, \boldsymbol{W} \sim \text{Normal}((W_1 - W_2 + 2\mathbb{1}_{\{1,2\}}(A)), 1)$ and the clinical outcome Y was simulated as $Y \mid S, A, W \sim \text{Bernoulli}(\text{expit}(-2+\mathbb{1}_{\{1,2\}}(A)+W_1/2-S/2))$. To simulate two-phase sampling, we allowed sampling probabilities to depend on vaccine A and outcome Y. In early-phase trials $P(\Delta=1 \mid A=a, Y=y)=1$ for all a, y, consistent with the standard design of measuring immunogenicity in all participants in

Table 3.2: **Details of generating scheme for each simulated trial set.** n is the sample size and $P_{y,a}$ is the sampling probability in the sub-population $Y = y$ and $A = a$

Scenario	Trial	Vaccine	n	$P(W_1)$	$P(W_2)$	$P_{1,0}$	$P_{1,1}$	$P_{0,0}$	$P_{0,1}$
1	1	1	200	0.65	0.80	1	1	1	1
	2	2	150	0.5	0.30	1	1	1	1
2	1	1	5000	0.65	0.80	0.05	0.1	0.05	0.1
	2	2	150	0.5	0.30	1	1	1	1
3	1	1	2000	0.65	0.80	0.05	0.1	0.05	0.1
	2	2	1500	0.5	0.30	0.05	0.1	0.05	0.1

Table 3.3: **Bias, variance, mean-squared error (MSE), coverage probability and width of 95% CI for first simulation**. Simulation results of three scenarios are summarized for two choices of referent populations. Our methods have consistent performance with small biases, low MSE and well-defined coverage probability of 95% confidence intervals. CI_c: CI coverage; CI_w: CI width.

Case	$\mathscr{T}_{\text{ref}}$	Vaccine	Truth	Bias	Variance	MSE	CI_c	CI_w
1	$\{1, 2\}$	1	2.0000	0.0058	0.0155	0.0171	0.9430	0.4820
1	$\{1, 2\}$	2	2.0000	-0.0068	0.0224	0.0218	0.9430	0.5779
1	$\{2\}$	1	1.8500	0.0009	0.0123	0.0122	0.9480	0.4341
1	$\{2\}$	2	1.8500	-0.0116	0.0371	0.0366	0.9390	0.7380
2	$\{1, 2\}$	1	1.8602	-0.0021	0.0041	0.0043	0.9450	0.2502
2	$\{1, 2\}$	2	1.8602	0.0021	0.0342	0.0314	0.9430	0.7078
2	$\{2\}$	1	1.8500	-0.0021	0.0041	0.0042	0.9460	0.2499
2	$\{2\}$	2	1.8500	0.0023	0.0354	0.0326	0.9370	0.7199
3	$\{1, 2\}$	1	2.0000	-0.0001	0.0140	0.0140	0.9450	0.4597
3	$\{1, 2\}$	2	2.0000	-0.0007	0.0206	0.0214	0.9230	0.5555
3	$\{2\}$	1	1.8500	-0.0008	0.0107	0.0102	0.9500	0.4034
3	$\{2\}$	2	1.8500	0.0001	0.0337	0.0349	0.9220	0.7058

such trials. For late-phase trials, we generated data with two-phase sampling according to probabilities listed in Table 3.2.

Our estimators exhibited low bias and reasonable MSE in all proposed scenarios (Table 3.3). They also achieve nominal confidence interval coverage with reasonable width of confidence intervals.

Table 3.4: **Details of generating scheme for the HVTN-inspired simulation.** N is the total sample size for each trial. P_y is the sampling probability given $Y = y$.

Trial	Vaccine	N	$\{P(W_j) : j = 1,\ldots,10\}$	P_1	P_0
1	1	1168	$\{0.546, 0.501, 0.443, 0.551, 0.537,$ $0.474, 0.409, 0.587, 0.468, 0.515\}$	0.57	0.06
2	2	200	$\{0.540, 0.432, 0.491, 0.444, 0.442,$ $0.503, 0.535, 0.532, 0.575, 0.576\}$	1	1
2	3	73	$\{0.534, 0.440, 0.420, 0.480, 0.498,$ $0.569, 0.506, 0.574, 0.432, 0.570\}$	1	1

3.3.2 Simulations inspired by HVTN trials

In the second simulation, we simulated three trials similar to the HVTN trials of interest: HVTN702 ($T = 1$), RV144 ($T = 2$) and HVTN097 ($T = 3$). For each trial, 10 binary covariates, W_j ($j = 1,\ldots,10$) were generated, $W_j \mid T \sim \text{Bernoulli}(p_j(T))$ with different generating probabilities p_j. For each trial t, the probabilities $p_j(t)$ were drawn at random from a Uniform$(0.4, 0.6)$ distribution; these probabilities are shown in Table 3.4. The sample sizes for each trial are fixed at the same level as the real trials ($n_1 = 1168, n_2 = 200, n_3 = 73$). In trial 1, 594 out of 1168 participants are observed in the active vaccine group, whereas for all the other three trials, only vaccinated subjects are documented and the vaccines used in HVTN702 ($A = 1$) is different than that used in either RV144 ($A = 2$) and in HVTN097 ($A = 3$). Next, we generated immune marker from Gaussian distribution, $S \mid A, \mathbf{W} \sim \text{Normal}(0.113\mathbb{1}_{>0}(A) + 0.025W_3 + 0.062W_5, 0.1683^2)$. The primary outcome Y in HVTN702 was simulated as $Y \mid S, A, \boldsymbol{W} \sim \text{Bernoulli}(\text{expit}(0.061S + 0.132\mathbb{1}_{>0}(A) + 0.086W_3 + 0.265W_5))$. A case-control sampling design was applied for $T = 1$, where $P(\Delta = 1 \mid Y = 1, A = 1, \boldsymbol{W}) = 0.57$ for all $\boldsymbol{W}$ and $P(\Delta = 1 \mid Y = 0, A = 1, \boldsymbol{W}) = 0.06$ for all $\boldsymbol{W}$. For each set of simulated trials, we estimated the average immunogenicity of each of the three vaccines standardized to the $T = 1$ population using our proposed estimators. This process was repeated 1000 times.

Our estimators again exhibited low bias, small MSE and well-calibrated 95% Wald-type confidence intervals in all scenarios (Table 3.5).

Table 3.5: **Results of HVTN inspired simulations in terms of bias, variance, mean-squared error (MSE), coverage probability and width of 95% CI**. Sample size reflects the number of participants in each trial. N: trial size; N_S: the number of participants having S measured; CI_c: CI coverage; CI_w: CI width.

$\mathscr{T}_{\text{ref}}$	Vaccine	N	N_S	Truth	Bias	Variance	MSE	CI_c	CI_w
1	1	1354	120	0.15735	0.00024	0.00025	0.00008	0.997	0.060
1	2	200	200	0.15735	0.00061	0.00017	0.00018	0.944	0.051
1	3	73	73	0.15735	-0.00181	0.00044	0.00043	0.947	0.081

3.4 Application to RV144 and HVTN Trials

The proposed methods were applied to three investigational HIV vaccine trials: HVTN702 [19], HVTN097 [18] and RV144 [34]. The vaccine regimen used in RV144 and HVTN097 is ALVAC-HIV-vCP1521 carrying clade 92TH023-AE, clade B gag, and clade B protease with adjuvant Alum, labeled as $P_{AE/B}$/alum; the vaccine regimen used in HVTN702 is ALVAC-vCP2438 carrying clade ZM96.C, clade B gag and clade B protease with adjuvant MF59, labeled as P_C/MF59. Our analysis characterized the immunogenicity of these vaccines in terms of their impact on CD4+ T cells expressing cytokines in response to ZM96. We evaluated the readout of this assay as both a continuous response magnitude (capped at 22000) and a binary response (0: Yes, 1: No), the latter indicating that the assay readout was greater than a positivity cutoff.

Baseline participant characteristics adjusted for in the analysis included age, sex, BMI, region of enrollment, and educational level.In HVTN702, the immune responses were measured subject to a case-control sampling scheme with known sampling weights. Participants who get vaccinated in any other two trials are naturally assigned weight one since the target immune markers are all measured.

We present results that compare the difference between vaccines evaluated in HVTN097 and RV144 ($P_{AE/B}$/alum) versus the vaccine evaluated in HVTN702 (P_C/MF59) standardized to the HVTN702 population. Results show for both naïve unadjusted estimator and our proposed TMLE estimator of mean difference in average response rate (RR) of CD4+ T cells expressing two cytokines (Table 3.6). The TMLE analysis indicated that

Table 3.6: **The difference in average immune responses of CD4+ cells between referent population HVTN702 and earlier trial population HVTN097 and RV144.** Comparisons were summarized for unadjusted approaches and our proposed method for both contrasts. ICS: Intracellular cytokine staining; RR: response rate; GM: geometric mean.

	HVTN702	HVTN097	RV144
Location	South Africa	South Africa	Thailand
Prevalent	Clade C	Clade C	Clade B
Vaccine	P_C/MF59	$P_{AE/B}$/alum	$P_{AE/B}$/alum
RR difference (unadj)	Ref	0.122 (-0.010, 0.254)	-0.207 (-0.316, -0.098)
p value	–	0.071	< 0.001
RR difference (TMLE)	Ref	0.170 (-0.026, 0.353)	-0.183 (-0.329, -0.028)
p value	–	0.088	0.021
GM ratio (unadj)	Ref	1.258 (0.945, 1.674)	0.728 (0.578, 0.918)
p value	–	0.116	0.007
GM ratio (TMLE)	Ref	1.455 (0.954, 2.218)	0.703 (0.519, 0.953)
p value	–	0.082	0.023

the HVTN702 vaccine had significantly higher RR and geometric mean values of the immune response when compared to RV144, with no evidence of difference in response rates comparing to the HVTN097 vaccine. The TMLE estimates were largely similar to the unadjusted estimates. These results may help explain the discrepant results between RV144 and HVTN702.

Chapter 4

Standardized Causal Effect Sizes in Vaccine Research

4.1 Introduction

The primary goal of this chapter is to explicitly discuss the considerations for defining standardized causal effect sizes across a wide range of scenarios. We also provide nonparametric efficient estimators accompanied by theory that elucidates the asymptotic properties.

The remainder of this chapter is structured as follows. In Section 4.2, we discuss the definition of effect sizes in a causal context in various scenarios. In Section 4.3 we first introduce one of novel standardized causal ES's, which presents an opportunity to differentiate the impact of effects beyond ATE. Next, we outline a detailed procedure of the estimation and inference for the proposed estimators. In Section 4.4 we assess finite sample performances of our estimators via extensive simulations.

4.2 Considerations for defining standardized causal effect sizes

4.2.1 Causal effects

We introduce the following general notation. We are interested in describing the impact of a binary intervention $A \in \{0,1\}$ on one or several outcomes. We denote a single outcome of interest by $Y \in \mathbb{R}$. We also assume that a set of contextual covariates $W \in \mathscr{W}$ is available, where $\mathscr{W}$ denotes the support of the random variable W. We use $O_1, \ldots, O_n$ to denote n independent and identically distributed copies of a random variable $O = (A, W, Y) \sim P \in \mathscr{M}$. Additionally, we define a counterfactual variable $Y(a)$, which corresponds to the outcome that would be observed if an individual were exposed to a specific intervention $a \in \{0,1\}$. We use P^a to denote the distribution of the counterfactual data unit $(W, A, Y(a))$ for $a = 0, 1$. Formulation of such potential outcomes generally requires the stable treatment value assumption (SUTVA), which stipulates that (i) there is only a single form of each intervention a and (ii) the potential outcomes for any given

individual do not depend on the interventions received by others. The additive ATE is defined as $E_{P^1}[Y(1)] - E_{P^0}[Y(0)]$, which is a common target for inference when evaluating an intervention. Here, we use the notation E_{P^a} to denote the expectation operator under distribution P^a. Other common targets for inference include the subgroup-specific ATE, e.g., $E_{P^1}[Y(1) \mid W = w] - E_{P^0}[Y(0) \mid W = w]$ for a specific covariate strata w and the average treatment effect amongst the treated $E_{P^1}[Y(1) \mid A = 1] - E_{P^0}[Y(0) \mid A = 1]$.

4.2.2 Standardized causal effect sizes

Similarly as with standardized effect sizes, we can define a standardized causal effect size by considering dividing the causal effect of interest by a relevant measure of variability of potential outcomes. In the most general form, we may consider estimands of the form

$$\psi = \frac{\mu(P^0, P^1)}{\sigma(P^0, P^1)},$$

where $\mu(P^0, P^1)$ is some causal contrast of interest summarizing a difference in the central tendency of the distributions of counterfactuals under intervention $A = 0$ versus $A = 1$ and $\sigma(P^0, P^1)$ is some measure of the spread of the counterfactuals under intervention $A = 0$ and/or $A = 1$. For simplicity, we restrict attention to additive causal contrasts so that $\mu(P^0, P^1)$ is assumed to be interpretable as an impact of the intervention in the units of the outcome. For example, if the ATE is of interest, then we may take $\mu(P^0, P^1) = E_{P^1}[Y(1)] - E_{P^0}[Y(0)]$. If instead, the ATT is of interest then we may take $\mu(P^0, P^1) = E_{P^1}[Y(1) \mid A = 1] - E_{P^0}[Y(0) \mid A = 1]$. However, it is not immediately clear what an appropriate choice of σ should be. In particular, we must clarify whether and how the choice of σ should vary by the choice of μ and the overarching scientific context of the experiment. In the following section, we discuss several scenarios and provide considerations for selection of σ in each case.

4.2.3 Choices of measure of counterfactual variability

Standardizing by the variability $Y(a)$. A natural choice for standardizing a causal effect is to select σ as the standard deviation of $Y(a)$, where a can be either 0 or 1. We use $\text{Var}^{1/2}(Y(a))$ to denote this quantity. The interpretation of the standardized causal effect size is thus as the mean difference in the outcomes presented in units of standard deviation of $Y(a)$. The choice of a in this context should be dictated by the scientific context depending on the desired interpretation. In many situations a control or a standard of care intervention is represented by one of the levels of a, say $a = 0$. In these situations, we suggest that standardization by $\text{Var}^{1/2}(Y(0))$ is a natural choice. For example, suppose we are interested in summarizing the effectiveness of a novel drug for cancer treatment relative to an existent and commonly-used therapy in terms of the therapies' impact on restricted mean survival time. It is important to note that the distribution of potential outcomes $Y(0)$ is closer to the real-world population, where the standard therapy is routinely given. Thus, the standardized causal effect size meaningfully captures the expected impact of adoption of the new therapy versus the standard therapy relative to the current variability of survival times.

Standardizing using a pooled counterfactual variance. More generally, we could take $\sigma(P^0, P^1) = [\beta \text{Var}(Y(0)) + (1-\beta)\text{Var}(Y(1))]^{1/2}$ for some fixed choice of weight $\beta \in [0,1]$. A natural choice is to set $\beta = P(A = 0)$, i.e., to weight counterfactual variances according to how frequently the associated intervention occurs in the natural world. In the context of a randomized trial, this estimand exactly corresponds to the estimand of Cohen's D statistic. However, this more general formulation allows for the possibility that intervention is confounded by other factors. We note that this form of σ can also be interpreted as $[\text{Var}(Y(A^*)) - \text{Var}(E[Y(A^*) \mid A^*])]^{1/2}$ where $A^* \sim \text{Bernoulli}(P(A = 1))$. That is, we can conceptualize a world in which the intervention $A = 1$ is given to exactly the same fraction of the population as in the observed world $P(A = 1)$; however, that fraction of the population is selected at random, rather than being influenced by confounding factors W.

Here is an example, we may be interested in evaluating the total cholesterol levels across different socioeconomic groups. It is not surprising that socioeconomic status (SES) can affect individuals' dietary habits, which in turn significantly influence their total cholesterol levels. However, real-world societies are diverse and always comprise multiple layers of socioeconomic patterns, which makes it unrealistic to assume all individuals are within one specific stratum. In this case, a meaningful choice of the measurement of variability would be a pooled variance based on sample proportions of subpopulations with each SES.

Standardizing by the variability $Y(a)$ in a subpopulation. If the chosen causal effect that constitutes the numerator of ψ is a subpopulation-specific effect (e.g., the ATT), then standardization by the variability of one of the potential outcomes in the entire population may not have a desirable interpretation. Instead, we may wish to align the choice of σ with the same subpopulation selected for μ. For example, we may consider $\sigma(P^0, P^1) = \text{Var}^{1/2}(Y(a) \mid A = 1)$ for either $a = 0$ or 1. In situations where the ATT is of primary interest, we suggest that choose $a = 1$ will often lead to the greatest interpretability. In this case ψ has an interpretation in terms of standardized units in the observed outcome since $\text{Var}^{1/2}(Y(1) \mid A = 1)$ equals the standard deviation of the observed outcome Y amongst individuals who naturally receive intervention level $A = 1$.

Special considerations for evaluating treatment effect heterogeneity using standardized causal effect sizes. If a fixed subpopulation defined by a particular covariate value w is of interest, then as above, we likely would wish to standardize the causal effect by the standard deviation of counterfactual outcomes in that subpopulation. Such a standardization would allow a natural comparison of, for example, the impact of an intervention in a subpopulation across several distinct outcomes. However, in studies that attempt to elucidate treatment effect heterogeneity, special consideration may be required. We generally say that treatment effect heterogeneity on the additive scale is present if the conditional ATE $E[Y(1) - Y(0) \mid W = w]$ is not constant in w. However, when comparing the magnitude of the impact of treatment across subpopulations defined by W is of interest, it may be more relevant in

some situations to define treatment effect heterogeneity on the standardized causal effect scale. That is, we could study

$$\psi(w) = \frac{E[Y(1) - Y(0) \mid W = w]}{\text{Var}^{1/2}(Y(a) \mid W = w)} ,$$

and ask whether $\psi(w)$ is constant in w. If the goal of an effect heterogeneity analysis is to describe whether the impact of an intervention varies by subgroup, then searching for effect modification on the scale of the ATE may mask relevant differences in subgroups if the variability of potential outcomes varies substantially across those subgroups. That is, the same ATE in two (or more) subgroups may yet reflect a meaningfully different real-world impact of the intervention in those subgroups.

The choice of causal effect sizes should be motivated by the specific context of studies. Different effect sizes may be appropriate depending on the nature of the research and underlying hypotheses. Moreover, it is crucial to choose a causal effect size that is interpretable and meaningful for real-world implications. The optimal option of causal effect size to report is not fixed. It varies depending on a combination of these factors, allowing us to offer distinct perspectives on the magnitude of a causal relationship.

4.3 Identification, estimation, and inference for standard causal effect sizes

Identification of standardized causal effect sizes must proceed on a case-by-case basis, depending on the choice of μ and σ. In this section, we present an in-depth analysis of one particular choice, with μ given by the ATE and σ equal to the standard deviation of $Y(0)$. We note that the variance of $Y(0)$ can be further decomposed into two components, $\text{Var}[Y(0)] = E[Y(0)^2] - E[Y(0)]^2$. We define $\psi_0 = E[Y(0)]$, $\psi_1 = E[Y(1)]$ and $\psi_2 = E[Y(0)^2]$, respectively. Hence, the ultimate causal estimand we are interested in iden-

tifying and estimating is

$$\psi = \frac{E[Y(1)] - E[Y(0)]}{\sqrt{E[Y(0)^2] - E[Y(0)]^2}} = \frac{\psi_1 - \psi_0}{\sqrt{\psi_2 - \psi_0^2}}.$$

There are multiple sets of assumptions under which we may be able to identify ψ_0, ψ_1, and ψ_2. The following set is sufficient for identification, while noting that other identifying assumptions exist.

Assumption 1 (Unconfoundedness): The assignment mechanism is independent of the potential outcomes conditional on observed covariates: $Y(a) \perp A \mid W$.

Assumption 2 (Positivity): For all possible values of w, the conditional probability of receiving intervention $A = 1$ is bounded between 0 and 1 $\delta < P(A = a \mid W) < 1 - \delta$ for some constant $\delta > 0$.

Essentially, *Assumption 1* implies that there are no unobserved confounders that affect both the treatment assignment and the outcomes. This assumption can be supported by careful study design and data collection strategies. Sensitivity analysis and graphical checks are often employed to assess the potential impact of violations. *Assumption 2* bounds the values of the propensity score away from zero. This assumption may be violated in observational studies due to contraindications to treatment. If both assumptions hold, then $E_{P^a}[Y(a)] = E[E(Y \mid A = a, W)]$ and $E_{P^a}[Y(a)^2] = E[E(Y^2 \mid A = a, W)]$, where we use E to denote expectation under the observed data distribution P. Thus, we have the following identification for ψ:

$$\psi = \frac{E[E(Y \mid A = 1, W)] - E[E(Y \mid A = 0, W)]}{\sqrt{E[E(Y^2 \mid A = 1, W)] - (E[E(Y \mid A = 0, W)])^2}}.$$

We use the short hand notation $\bar{Q}(A, W) = E(Y \mid A, W)$ and $\bar{Q}_2(A, W) = E(Y^2 \mid A, W)$ to denote relevant components of this identification result.

4.3.1 Efficiency theory

It is straightforward to show that the efficient influence function in a nonparametric model for a parameter of the form $E[E(f(Y) \mid A = a, W)]$ for a some non-degenerate transformation f of Y is

$$D_f(O_i) = \frac{I(A_i = a)}{P(A = a \mid W = W_i)} \{Y_i - E[f(Y) \mid A = a, W]\} + E[f(Y) \mid A = a, W] - E\{E[f(Y) \mid A = a, W]\} .$$

This result can be directly applied to establish the EIF for ψ_0, ψ_1, and ψ_2. The delta method then implies that the EIF for ψ is:

$$\begin{aligned} D(O_i) = & \frac{\psi_1 \psi_0 - \psi_2}{(\psi_2 - \psi_0^2)^{3/2}} \frac{I(A_i = 0)}{P(A = 0 \mid W = W_i)} \left[Y_i - \bar{Q}(0, W_i)\right] \\ & + \frac{1}{(\psi_2 - \psi_0^2)^{1/2}} \frac{I(A_i = 1)}{P(A = 1 \mid W = W_i)} \left[Y_i - \bar{Q}(1, W_i)\right] \\ & - \frac{\psi_1 - \psi_0}{2(\psi_2 - \psi_0^2)^{3/2}} \frac{I(A_i = 0)}{P(A = 0 \mid W = W_i)} \left[Y_i^2 - \bar{Q}_2(0, W_i)\right] \\ & + \frac{\psi_1 \psi_0 - \psi_2}{(\psi_2 - \psi_0^2)^{3/2}} \left[\bar{Q}(0, W_i) - \psi_0\right] \\ & + \frac{1}{(\psi_2 - \psi_0^2)^{1/2}} \left[\bar{Q}(1, W_i) - \psi_1\right] \\ & + \frac{\psi_1 - \psi_0}{2(\psi_2 - \psi_0^2)^{3/2}} \left[\bar{Q}_2(0, W_i) - \psi_2\right] \end{aligned}$$

4.3.2 Plug-in estimation

We propose a plug-in estimator of ψ that is based on estimators of (i) the conditional density of Y given A and W and (ii) the marginal distribution of W. We denote the estimated conditional density of Y evaluated at y given $A = a$ and $W = w$ by $q_n(y \mid A = a, W = w)$. We note that given such a conditional density estimate, estimates $\bar{Q}_n(a, w)$ and $\bar{Q}_{n,2}(a, w)$

of $\bar{Q}(a,w)$ and $\bar{Q}_2(a,w)$, respectively are implied for any given values (a,w):

$$\bar{Q}_n(a,w) = \int y\, q_n(y \mid a,w)dy \tag{4.1}$$

$$\bar{Q}_{n,2}(a,w) = \int y^2\, q_n(y \mid a,w)dy\ . \tag{4.2}$$

We propose to use the empirical distribution of W to estimate its marginal distribution. Thus, a plug-in estimator may be computed as

$$\psi_n = \frac{\frac{1}{n}\sum_{i=1}^{n}[\bar{Q}_n(1,W_i) - \bar{Q}_n(0,W_i)]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\bar{Q}_{n,2}(0,W_i) - [\frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n(0,W_i)]^2}}\ . \tag{4.3}$$

4.3.3 One-step corrected estimation

If flexible estimators of the conditional density of Y are used in construction of the plug-in estimator then it will have non-standard asymptotic behavior. In particular, we expect that the plug-in estimator will have large finite-sample bias that will not shrink to zero at an appropriate rate. To facilitate the usage of flexible density estimators, we consider a one-step correction [6]. This estimator can be implemented in the following steps.

Estimate the probability of receiving intervention a given covariates. Use either parametric (e.g., logistic regression) or a nonparametric regression (e.g., kernel regression) to regress the A onto W. Based on the model fit, evaluate the estimated probability that $A = a$ given $W = W_i$ for $i = 1,\ldots,n$. Denote this estimate by $g_n(a \mid W_i)$.

Estimate the conditional density of outcome given intervention and covariates. Use a conditional density estimation technique (e.g., kernel density estimation) to obtain an estimate q_n of the conditional density of Y given A,W.

Use numeric integration to obtain estimates of conditional means. Evaluate (4.1) for $a = 0,1$ and for $w = W_1,\ldots,W_n$ to obtain $\bar{Q}_n(a,W_i)$ for $i = 1,\ldots,n$. Evaluate (4.2) for $a = 0$ and $w = W_1,\ldots,W_n$ to obtain $\bar{Q}_{n,2}(0,W_i)$ for $i = 1,\ldots,n$.

Construct plug-in estimate of standardized causal effect size. Evaluate plug-in estima-

tor in (4.3) based on estimates obtained in previous step.

Evaluate efficient influence function using estimated quantities. Evaluate

$$\begin{aligned} D_n(O_i) = & \frac{\psi_{n,1}\psi_{n,0} - \psi_{n,2}}{(\psi_{n,2} - \psi_{n,0}^2)^{3/2}} \frac{I(A_i = 0)}{g_n(0 \mid W_i)} \left[Y_i - \bar{Q}_n(0, W_i)\right] \\ & + \frac{1}{(\psi_{n,2} - \psi_{n,0}^2)^{1/2}} \frac{I(A_i = 1)}{g_n(1 \mid W_i)} \left[Y_i - \bar{Q}_n(1, W_i)\right] \\ & - \frac{\psi_{n,1} - \psi_{n,0}}{2(\psi_{n,2} - \psi_{n,0}^2)^{3/2}} \frac{I(A_i = 0)}{g_n(0 \mid W_i)} \left[Y_i^2 - \bar{Q}_{n,2}(0, W_i)\right] \\ & + \frac{\psi_{n,1}\psi_{n,0} - \psi_{n,2}}{(\psi_{n,2} - \psi_{n,0}^2)^{3/2}} \left[\bar{Q}_n(0, W_i) - \psi_{n,0}\right] \\ & + \frac{1}{(\psi_{n,2} - \psi_{n,0}^2)^{1/2}} \left[\bar{Q}_n(1, W_i) - \psi_{n,1}\right] \\ & + \frac{\psi_{n,1} - \psi_{n,0}}{2(\psi_{n,2} - \psi_{n,0}^2)^{3/2}} \left[\bar{Q}_{n,2}(0, W_i) - \psi_{n,2}\right] \ , \end{aligned}$$

for $i = 1, \ldots, n$, where $\psi_{n,0} = n^{-1} \sum_{i=1}^{n} \bar{Q}_n(0, W_i)$, $\psi_{n,1} = n^{-1} \sum_{i=1}^{n} \bar{Q}_n(1, W_i)$, and $\psi_{n,2} = n^{-1} \sum_{i=1}^{n} \bar{Q}_{n,2}(0, W_i)$.

Compute the final one-step estimate of standardized causal effect size. The one-step estimate is defined as $\psi_n^+ = \psi_n + n^{-1} \sum_{i=1}^{n} D_n(O_i)$.

4.3.4 Asymptotic study of one-step estimator

We have the following theorem pertaining to the one-step estimator. We use $||f_n|| = \int f_n(o) dP(o)$ to denote the $L^2(P)$ norm of a given P-measurable function f_n. We adopt the shorthand $g(a \mid W) = P(A = a \mid W)$.

Theorem 5. *Suppose that*

(A1) $||g_n - g|| \, ||\bar{Q}_n - \bar{Q}|| = o_P(n^{-1/2})$

(A2) $||g_n(0 \mid \cdot) - g(0 \mid \cdot)|| \, ||\bar{Q}_{n,2} - \bar{Q}_2|| = o_P(n^{-1/2})$

(A3) $||D_n - D||^2 = o_P(1)$ *and* D_n *falls in a P-Donsker class with probability tending to 1 as n tends to infinity*

Then $\psi_n^+ = \psi + n^{-1}\sum_{i=1}^n D(O_i) + o_P(n^{-1/2})$.

The proof of the theorem follows using standard analysis of one-step estimators. An immediate implication of Theorem 5 is that $n^{1/2}\psi_n^+$ has a Normal limiting distribution with mean ψ and variance equal to the variance of the random variable $D(O)$. This asymptotic variance can be consistently estimated by $n^{-1}\sum_{i=1}^n [D_n(O_i) - n^{-1}\sum_{j=1}^n D_n(O_j)]^2$. Another implication of the theorem is double robustness of the one-step estimator, which states that if either (i) $||g_n - g|| = o_P(1)$ or (ii) both $||\bar{Q}_n - \bar{Q}|| = o_P(1)$ and $||\bar{Q}_{2,n}(0,\cdot) - \bar{Q}(0,\cdot)|| = o_P(1)$, then $\psi_n^+ - \psi = o_P(1)$.

4.4 Simulation study

We investigated the finite-sample performance of plug-in and one-step corrected estimators of the proposed estimand described in Section 4.3. For this study, 1000 datasets were generated at four sample sizes $n \in \{250, 500, 1000, 2000\}$ under the following data-generating mechanism. One normally distributed covariate $W \overset{iid}{\sim} N(3,1)$ was generated. Given W the intervention A was simulated as a Bernoulli random variable with $P(A = 1 \mid W) = \text{expit}(1.2 - 0.3W)$. Given A and W, the outcome Y was simluated as $Y \mid A, W \overset{iid}{\sim} N(-1.3W + 2A, 1 + 2A)$. Based on this data generating process, the true value of $\psi = 1.2187$.

We considered constructing our estimates in two different ways. The first used maximum likelihood estimation to estimate the density of Y given A, W; the second used nonparametric kernel methods for density estimation. Leave-one-out cross-validation was used to select the bandwidth of the kernel density estimation method. All estimators are assessed in terms of bias, variance, mean squared error (MSE) and relative efficiency ($\text{RE} = \text{MSE}_{\text{one-step}}/\text{MSE}_{\text{plug-in}}$). $RE < 1$ indicates the one-step estimator has better finite sample efficiency than plug-in estimator.

Results are reported in Table 4.1. The one-step estimator outperformed the plug-in estimator in most of the scenarios that we evaluated. In the smallest sample size, the plug-

in estimator enjoyed minor improvements in MSE relative to the one-step, owing primarily to reduced bias. However, as the sample size increased, the relative performance of the one-step improved. When sample size is greater than 500, efficiency gains of the one-step estimator were between 3.66 and 5.39 percent. As expected based on theory, we found that the plug-in estimator when coupled with nonparametric kernel density estimation exhibited large bias that decreased to zero slowly and that, when scaled by $n^{1/2}$, diverges. On the other hand, the one-step correction appropriately removes bias from the plug-in estimate and recovers standard $n^{1/2}$-asymptotic behavior.

Type	Method	n	Estimates	Bias	Variance	MSE	RE
parametric	plug-in	250	1.2221	0.0035	0.0376	0.0376	1.0000
parametric	plug-in	500	1.2191	0.0004	0.0172	0.0172	1.0000
parametric	plug-in	1000	1.2190	0.0004	0.0083	0.0083	1.0000
parametric	plug-in	2000	1.2192	0.0006	0.0043	0.0042	1.0000
parametric	CDE	250	1.2293	0.0106	0.0380	0.0381	1.0146
parametric	CDE	500	1.2225	0.0039	0.0174	0.0174	1.0132
parametric	CDE	1000	1.2207	0.0020	0.0083	0.0083	1.0046
parametric	CDE	2000	1.2201	0.0015	0.0043	0.0043	1.0003
nonparametric	plug-in	250	1.1612	-0.0574	0.0332	0.0365	1.0000
nonparametric	plug-in	500	1.1863	-0.0323	0.0171	0.0181	1.0000
nonparametric	plug-in	1000	1.1992	-0.0194	0.0084	0.0088	1.0000
nonparametric	plug-in	2000	1.2074	-0.0112	0.0043	0.0045	1.0000
nonparametric	CDE	250	1.2287	0.0101	0.0371	0.0372	1.0181
nonparametric	CDE	500	1.2239	0.0052	0.0171	0.0171	0.9461
nonparametric	CDE	1000	1.2224	0.0038	0.0085	0.0085	0.9634
nonparametric	CDE	2000	1.2212	0.0026	0.0043	0.0043	0.9616

Table 4.1: Simulation study: bias, variance, mean squared error (MSE) and relative efficiency (RE) of one-step and plug-in estimators over 1000 Monte Carlo simulations ($n \in \{250, 500, 1000, 2000\}$).

4.5 Discussion

In this study, we present a framework for defining and standardizing effect sizes in a causal context. We address the challenges associated with evaluating causal effects across different outcome scales and studies, and propose a fresh perspective that enables fair com-

parisons. Particularly, we go beyond relying solely on difference-based causal estimands by standardizing them with respect to the standard deviation of counterfactuals within a specific population of interest. The choice of such population depends on the research objectives. Additionally, we introduce nonparametric estimators for causal effect sizes, providing a rigorous methodology for their estimation. We believe that our approach offers a promising tool for researchers and practitioners seeking to assess causal effects in a robust and consistent manner. One open challenge lies in the potential application of targeted learning to enhance estimation techniques and integrate modern machine learning methods for estimating high dimensional nuisance parameters.

Appendix A

Appendix for Chapter 3

A.1 Proof of Theorem 1

Proof. If (A1) and (A2) hold then

$$
\begin{aligned}
\psi_{\mathbb{X}}(a) &= E_{\mathbb{X}}[S(a) \mid T = \mathscr{T}_{\text{ref}}] \\
&= E_{\mathbb{X}}\left\{E_{\mathbb{X}}\left[S(a) \mid T \in \mathscr{T}_{\text{ref}}, \boldsymbol{W}_S\right)\right] \mid T \in \mathscr{T}_{\text{ref}}\} && \text{tower rule} \\
&= E_{\mathbb{X}}\left\{E_{\mathbb{X}}\left[S(a) \mid \boldsymbol{W}_S\right] \mid T \in \mathscr{T}_{\text{ref}}\right\} && S(a) \perp T \mid \boldsymbol{W}_S \\
&= E_{\mathbb{X}}\left\{E_{\mathbb{X}}\left[S(a) \mid A = a, \boldsymbol{W}_S\right] \mid T \in \mathscr{T}_{\text{ref}}\right\} && S(a) \perp A \mid \boldsymbol{W}_S \\
&= E_X\left\{E_X\left(S \mid A = a, \boldsymbol{W}_S\right) \mid T \in \mathscr{T}_{\text{ref}}\right\} && \text{consistency}
\end{aligned}
$$

□

Bibliography

[1] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

[2] Peter C Austin and Elizabeth A Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28): 3661–3679, 2015.

[3] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[4] Angelika Banzhoff, Pantaleo Nacci, and Audino Podda. A new MF59-adjuvanted influenza vaccine enhances the immune response in the elderly with chronic diseases: results from an immunogenicity meta-analysis. *Gerontology*, 49(3):177–184, 2003.

[5] Linda-Gail Bekker, Zoe Moodie, Nicole Grunenberg, Fatima Laher, Georgia D Tomaras, Kristen W Cohen, Mary Allen, Mookho Malahleha, Kathryn Mngadi, Brodie Daniels, et al. Subtype C ALVAC-HIV and bivalent subtype C gp120/MF59 HIV-1 vaccine in low-risk, HIV-uninfected, South African adults: a phase 1/2 trial. *The lancet HIV*, 5(7):e366–e378, 2018.

[6] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

[7] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[8] Patrizia Cavazzoni. Coronavirus (COVID-19) Update: FDA Limits Use of Certain Monoclonal Antibodies to Treat COVID-19 Due to the Omicron Variant, Jan 2022. URL https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-limits-use-certain-monoclonal-antibodies-trea

[9] Centers for Disease Control and Prevention. Diagnoses of hiv infection in the united states and dependent areas, 2019. http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html, 2021. Accessed [June 21, 2021].

[10] Zhenghao Chu. Using deep learning methods to predict the vrc01 neutralization sensitivity by hiv-1 gp160 sequence features. Master's thesis, Emory University, 2020. URL https://etd.library.emory.edu/concern/etds/q237ht29n?locale=en.

[11] Amy W Chung, Musie Ghebremichael, Hannah Robinson, Eric Brown, Ickwon Choi, Sophie Lane, Anne-Sophie Dugast, Matthew K Schoen, Morgane Rolland, Todd J Suscovich, et al. Polyfunctional Fc-effector profiles mediated by IgG subclass selection distinguish RV144 and VAX003 vaccines. *Science translational medicine*, 6 (228):228ra38–228ra38, 2014.

[12] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[13] Lawrence Corey, John R Mascola, Anthony S Fauci, and Francis S Collins. A strategic approach to covid-19 vaccine r&d. *Science*, 368(6494):948–950, 2020.

[14] Lawrence Corey, Peter B Gilbert, Michal Juraska, David C Montefiori, Lynn Morris, Shelly T Karuna, Srilatha Edupuganti, Nyaradzo M Mgodi, Allan C deCamp, Erika Rudnicki, et al. Two randomized trials of neutralizing antibodies to prevent hiv-1 acquisition. *New England Journal of Medicine*, 384(11):1003–1014, 2021.

[15] Luis Furuya-Kanamori, Chang Xu, Suhail AR Doi, Justin Clark, Kinley Wangdi, Deborah J Mills, and Colleen L Lau. Comparison of immunogenicity and safety of licensed Japanese encephalitis vaccines: A systematic review and network meta-analysis. *Vaccine*, 39(32):4429–4436, 2021.

[16] Peter Gilbert, Steve Self, Malla Rao, Abdollah Naficy, and John Clemens. Sieve analysis: methods for assessing from vaccine trial data how vaccine efficacy varies with genotypic and phenotypic pathogen variation. *Journal of clinical epidemiology*, 54(1):68–85, 2001.

[17] Gene V Glass. Primary, secondary, and meta-analysis of research. *Educational researcher*, 5(10):3–8, 1976.

[18] Glenda E Gray, Ying Huang, Nicole Grunenberg, Fatima Laher, Surita Roux, Erica Andersen-Nissen, Stephen C De Rosa, Britta Flach, April K Randhawa, Ryan Jensen, et al. Immune correlates of the Thai RV144 HIV vaccine regimen in South Africa. *Science translational medicine*, 11(510):eaax1880, 2019.

[19] Glenda E Gray, Linda-Gail Bekker, Fatima Laher, Mookho Malahleha, Mary Allen, Zoe Moodie, Nicole Grunenberg, Yunda Huang, Doug Grove, Brittany Prigmore, et al. Vaccine efficacy of ALVAC-HIV and bivalent subtype C gp120–MF59 in adults. *New England Journal of Medicine*, 384(12):1089–1100, 2021.

[20] Robert J Grissom and John J Kim. *Effect sizes for research: A broad practical approach.* Lawrence Erlbaum Associates Publishers, 2005.

[21] Susan Gruber and Mark J van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1), 2010.

[22] Larry V Hedges. Distribution theory for glass's estimator of effect size and related estimators. *journal of Educational Statistics*, 6(2):107–128, 1981.

[23] Nima S Hejazi, Mark J van der Laan, Holly E Janes, Peter B Gilbert, and David C Benkeser. Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials. *Biometrics*, 77 (4):1241–1253, 2021.

[24] Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

[25] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[26] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[27] Guanglei Hong and Stephen W Raudenbush. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101(475):901–910, 2006.

[28] Silke Janitza, Ender Celik, and Anne-Laure Boulesteix. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, 12(4):885–915, 2018.

[29] Cheng Ju, Susan Gruber, Samuel D Lendle, Antoine Chambaz, Jessica M Franklin, Richard Wyss, Sebastian Schneeweiss, and Mark J van der Laan. Scalable collaborative targeted learning for high-dimensional data. *Statistical methods in medical research*, 28(2):532–554, 2019.

[30] Craig A Magaret, David C Benkeser, Brian D Williamson, Bhavesh R Borate, Lindsay N Carpp, Ivelin S Georgiev, Ian Setliff, Adam S Dingens, Noah Simon, Marco Carone, et al. Prediction of vrc01 neutralization sensitivity by hiv-1 gp160 sequence features. *PLoS computational biology*, 15(4):e1006952, 2019.

[31] Lynn Morris and Nonhlanhla N Mkhize. Prospects for passive immunity to prevent hiv infection. *PLoS medicine*, 14(11):e1002436, 2017.

[32] Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

[33] Maia A Rabaa, Yves Girerd-Chambaz, Kien Duong Thi Hue, Trung Vu Tuan, Bridget Wills, Matthew Bonaparte, Diane Van Der Vliet, Edith Langevin, Margarita Cortes, Betzana Zambrano, et al. Genetic epidemiology of dengue viruses in phase III trials of the CYD tetravalent dengue vaccine and implications for efficacy. *Elife*, 6:e24196, 2017.

[34] Supachai Rerks-Ngarm, Punnee Pitisuttithum, Sorachai Nitayaphan, Jaranit Kaewkungwal, Joseph Chiu, Robert Paris, Nakorn Premsri, Chawetsan Namwat, Mark de Souza, Elizabeth Adams, et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *New England Journal of Medicine*, 361(23): 2209–2220, 2009.

[35] James M Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN, 2000.

[36] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[37] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.

[38] Marcella Sarzotti-Kelsoe, Robert T Bailer, Ellen Turk, Chen-li Lin, Miroslawa Bilska, Kelli M Greene, Hongmei Gao, Christopher A Todd, Daniel A Ozaki, Michael S Sea-

man, et al. Optimization and validation of the tzm-bl assay for standardized assessments of neutralizing antibodies against hiv-1. *Journal of immunological methods*, 409:131–146, 2014.

[39] Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

[40] Mark J van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

[41] Mark J van der Laan and Susan Gruber. Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics*, 6(1), 2010.

[42] Mark J van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

[43] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[44] Mark J van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

[45] Hyejin Yoon, Jennifer Macke, Anthony P West Jr, Brian Foley, Pamela J Bjorkman, Bette Korber, and Karina Yusim. Catnap: a tool to compile, analyze and tally neutralizing antibody panels. *Nucleic acids research*, 43(W1):W213–W219, 2015.