Quantifying and mitigating selection bias in probability and nonprobability samples

Valerie C Bradley

New College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2023

Acknowledgements

First and foremost, an enormous thank you to my supervisors Seth Flaxman and Dino Sejdinovic for your endless patience and encouragement over the last five years. Thank you for bearing with my (perhaps too) frequent forays back into campaign world or new continents, and for understanding that my passion for this work come from application. Thank you in particular for herding me over the finish line, I would not have made it across without you.

I am endlessly grateful to my collaborators. Thank you to Tom Nichols for your enthusiasm for and guidance of my first mini project. To Xiao-Li Meng, it has been a privilege to work with and learn from you. Thank you to Shiro Kuriwaki and Michael Isakov for seemingly endless Zooms and revisions. Liza Semenova, you are a joy to work with, and thank you for sharing your home with me. A special thank you to Meg Schwenzfeier who is not only a collaborator here, but someone who understands the academic and applied halves of my work and my brain better than anyone.

I firmly believe that Oxford is the best place on the planet to be a graduate student. Not only because of the peerless faculty and academic opportunities it offers, but also due to the opportunities it provides to build community, in your department and out. Oxford has given me some of my best friends in the world, for which I am eternally grateful. Catherine – you and 18 Jericho St will always be home to me. To my ppl (Camilla, Sarah, and Alina), fellow HFA alum (Kate), seafood aficionados (Tyler, Josh, Michael, Will, and Brendan), women's volleyball co-captain (KG), Aussie MPPs (Chisso and Jim) – you all mean the world to me, and I can't wait to cook you pasta soon.

Thank you to OxWaSP, and my cohort in particular – Ana, James, Natalia, Bobby, Hector, Lorenzo, Alan, Deborah, Will, Lucy, and Maude. It was a joy to work with you, and thank you for indulging my love of politics and polling. Thank you to the entire Department of Statistics for all the camaraderie, support, and laughter over the years. This research would not have been possible without the support of the Clarendon Scholarship and New College.

Lastly, to my family. Anna – thank you for being the biggest fan of every season of the soap opera, I can't wait to keep watching with you. Scott – you are wise and brilliant and passionate, and thank you for sharing that with us. Charlie – somehow you're both still my tiny cherub and my favorite partner in crime. Mom

and Dad, in so many ways I would not be here without you. You are unfathomably generous and ferociously and unconditionally (as long as I finish my PhD) loving. You are my safety net and my wings. I love you all so much.

Statement of Originality

I hereby declare that except where specific reference is made to the work of others, the content of this thesis is my own work and has not been submitted in whole or in parts for any other degree or qualification. This thesis is my own work unless otherwise stated in the authorship form at the end of the chapters.

Valerie C Bradley
Michaelmas 2023

Contents

List of Abbreviations

AAPOR . . . American Association of Public Opinion Researchers

ACS . . . . . American Community Survey

AL . . . . . Active Learning

ALSD . . . . . Active Learning Sampling Design

ASD . . . . . Adaptive Survey Design

BART . . . . . Bayesian Additive Regression Tree

BO . . . . . Bayesian Optimization

CDC . . . . . US Centers for Disease Control and Prevention

DAG . . . . . Directed acyclic graph

ddc . . . . . data defect correlation

ddi . . . . . data defect index

deff . . . . . Kish's design effect

HSE . . . . . Health Survey for England

IPW . . . . . Inverse probability weighting

MAR . . . . . Missing at random

MCAR . . . . Missing completely at random

MNAR . . . . Missing not at random

MRP . . . . . Multi-level regression and post-stratification

mVAM . . . . Mobile Vulnerability Analysis and Mapping

NHS . . . . . National Health Service (UK)

SCM . . . . . Structural causal model

SRS . . . . . Simple random sample

TSE . . . . . Total Survey Error

WFP . . . . . World Food Programme

1 Introduction

This thesis is an integrated thesis and contains 6 chapters in total, including 4 original research papers, an introduction, and a conclusion. Each research paper chapter is intended to be stand-alone, so will itself contain an introduction, literature review, and conclusion. Therefore the purpose of this introduction chapter is to outline the overarching motivation for my research and place each paper in the broader context of the challenges faced by modern survey research.

I start with the motivation for this work, then introduce fundamental concepts underlying survey research, including probability sampling, selection bias, and methods for quantifying and adjusting for selection bias. Lastly, I outline the structure of the remaining chapters, and describe how each piece of work contributes to the discourse on the prevention of, quantification of, and recovery from selection bias.

1.1 Motivation

The field of survey research has undergone dramatic transformation in the last decade. Historically, survey research has relied primarily on probability samples, and the strong mathematical framework accompanying them, to ensure and evaluate quality of research. Probability samples are defined by their random selection mechanism – each unit in the population has a non-zero probability of being observed that is known at the time of design. Probability surveys thrived in the 1960s-1990s, when landline telephones were ubiquitous, response rates and institutional trust were high, and telephone and face-to-face surveys were the primary tools that social scientists had for learning about populations (Groves, 2011).

Since the 1990s, the conditions that enabled this "golden era of survey research" have deteriorated. Firstly, response rates have declined precipitously in the last 30 years. For example, in the 1990s survey researchers were able to conduct surveys by calling landline telephones almost exclusively, however in the last 30 years the proportion of American adults with landlines has plummeted. In January 2003, the Centers for Disease Control and Prevention (CDC), the US national public health agency, estimated that over 95% of American adults live in households with a landline telephone (Blumberg and Luke, 2007), but as of the CDC's latest estimates from July-December of 2022, only 17% of American adults live in a household with a landline (Blumberg and Luke, 2022). American adults are instead largely wireless-only (as of Dec 2022, over 80% of American adults live in cell-only households), but cell phone numbers are more difficult to use in survey research than landlines because they suffer from higher turnover, are more transient, and for the most part must be acquired through commercial data vendors rather than government sources like voter registration forms.

Second, Kennedy and Hartig (2019) find that response rates to telephone surveys declined by 31 percentage points from 1997 to 2018; in 1997, 37% of people attempted in Pew Research's telephone surveys responded, but by 2018, only 6% of people responded to telephone surveys. This problem is not limited to landline surveys – from 1997-2006 Pew only contacted landline numbers, however the decline in response rates continued from 2006 to 2018 after cell phones were introduced.

Third, institutional trust has declined in the last 3 decades. Gallup, a polling organization, tracks trust in a range of public institutions and has found consistent, significant declines in trust for a wide range of US institutions, including television news and newspapers, and current trust is at or near their all-time lows since Gallup started tracking in 1979 (Saad, 2023). This low institutional trust likely contributes to the public's declining willingness to participate in surveys on any mode – landline, cell, or otherwise.

Simultaneously, the rise of the internet and other tools for collecting cheap, large-scale data have paved the way for the rise of nonprobability, opt-in online surveys. Nonprobability samples include all those that do not fit the strict definition of probability surveys, either because some units in the population have probability equal to 0 of being observed (for example because they are not members of an online panel), or if individual behavior, rather than sampling design, governs the probability of response.

The environment has been changing for decades, but survey researchers have only adapted more recently. Many well-respected, large-scale surveys only started

including an online mode in the past 10 years. For example, Pew Research added an online component to their American Trends Panel in 2014 (Keeter, 2019), Gallup added an online component to their Well-Being Index in 2018 (Gallup, 2018), and in the UK, the Understanding Society panel tested an online mode in 2015, but did not incorporate it fully until 2016 (Carpenter, 2018). YouGov is an online pollster that is now highly-respected for robust methodology, but as recently as 2014, when the New York Times and CBS began to use YouGov in their pre-election polling, the American Association of Public Opinion Research (AAPOR) released a statement condemning the use of an opt-in nonprobability panel (Link, 2014).

This fracturing contactability landscape has culminated in a series of catastrophic misses for the survey research industry in the last decade. In 2016, public opinion polling failed to predict that the UK would vote to leave the EU and that Donald Trump would be elected president of the United States (Jackson, 2016; Kennedy et al., 2018). Despite much analysis and debate following these two high-profile polling failures and strong performance in the 2018 US midterms, public polls mispredicted the outcome of the 2019 Australian federal election (Pennay et al., 2020b), and again overestimated support for Democrats in the 2020 US presidential and congressional elections (Clinton et al., 2020).

The debate between probability and nonprobability samples has raged over the last decade, with each new election seeming to produce a different conclusion about which approach should prevail. However this debate strikes me as misguided. As I will argue in the rest of this chapter, there is no such thing as a true probability sample in modern survey research. The surveys commonly referred to as "probability surveys" are those that randomly select a list of potential respondents from some type of (we assume) complete sampling frame, however the probabilistic component of a "probability" sample is one step in a series of data collection steps, most of which are non-probabilistic. I will explain, using the framework for decomposing error in estimates of population means by Meng (2018), how probabilistic selection stages do not efficiently counteract selection bias introduced at other stages of the data collection process.

Consider random digit dialing (RDD), for example, which was once seen as a gold-standard probabilistic selection mechanism. RDD randomly generates telephone numbers to attempt to survey, so theoretically all phone numbers have some non-zero probability of selection. However, in order to do any sort of sub-national geographic targeting using RDD, you either must rely on phone number area codes (therefore systematically excluding more transient populations), or cast a wide net and ask

respondents to self-report geography, screening out people who fall outside of the desired population. The latter is highly inefficient and likely prohibitively expensive.

The relevant question is not *whether* selection bias exists in a survey, but rather *how much* exists and the degree to which it is possible to recover unbiased estimators. This view is very recently beginning to emerge in the survey research field, both in academic papers (Bailey, 2023) and in discussion of pre-election polling of the 2022 US midterm elections (Rutenberg et al., 2022), which saw an interesting dichotomy of less-frequent polling from higher-quality pollsters (that were highly accurate) and a flood of surveys from less reputable pollsters (that were very inaccurate).

Polling is not fundamentally broken as a tool for understanding populations, but accurate polling requires relentless dedication to combating selection bias from every possible angle. There are, of course, other sources of error in surveys, as outlined by the Total Survey Error (TSE) framework (Biemer and Lyberg, 2003). However here I focus solely on selection bias, which impacts a number of types of error outlined in the TSE framework, including coverage error, nonresponse error, and processing error.

My work in this thesis aims to tackle selection bias at every step of the data collection process by improving the set of tools at a researcher's disposal for preventing, quantifying, addressing, and communicating the impact of selection bias. Chapter 2 demonstrates the utility of the data defect correlation (*ddc*) framework for quantifying selection bias, and Chapter 3 builds on this framework, as well as Bayesian optimization and active learning, to introduce a sampling framework that attempts to predict and preempt selection bias that results from unit nonresponse. Chapter 4 introduces *leverage* as a metric for quantifying population frame uncertainty, an under-appreciated source of error in adjustment for nonresponse, and develops tools for communicating that uncertainty. Lastly, Chapter 5 examines the performance of common nonresponse adjustment methods, and one novel method, when applied to a large nonprobability survey.

1.2 Background

As this thesis follows an integrated format, each chapter in the thesis will be self-contained and have its own background section explaining technical details required to understand its contributions. In this section, therefore, I aim to set the stage of modern survey research more broadly and argue why better approaches for tackling selection bias are desperately needed. I start by outlining probability sampling, the foundation upon which survey research is built. I then introduce the problem of selection bias, and discuss the standard set of tools used by survey researchers

to combat it. Next, I discuss nonprobability samples, and methods for evaluating their quality. Lastly, I discuss the challenges faced by probability samples and recovering from selection bias in practical survey research.

1.2.1 Probability samples

A *probability sample* is a randomly selected sample in which the probability that a unit appears in a sample is known and strictly positive for all units in the population (Lohr, 2010). The simplest type of probability sample is the simple random sample (SRS). For a finite population of units $i = 1, \ldots, N$, let $S_i = 1$ if the ith unit is selected into the sample, and 0 otherwise. An SRS is a probability sample without replacement in which $\mathbb{P}(S_i = 1) \propto 1/N$ for all units in the population.

A probability sample may have unequal probabilities of selection, as long as those probabilities are 1) known and 2) strictly positive (such that no population units are systematically excluded from the sample). For example, stratified random sampling is a probabilistic sampling technique in which the population is divided into strata based on some auxiliary features observed for the population, which we will denote $\mathbf{X}$. Then, units are selected using SRS within each stratum, rather than from the entire population, which helps reduce the sampling variance relative to SRS (Lohr, 2010). In stratified random sampling, the probability of selection may be proportional to stratum size, or may be unequally weighted such that some strata are oversampled relative to population size. However as long as the probability of selection is known in advance and is non-zero for all units, the sample is still a valid probability sample.

In a probability sample, the sampling mechanism is identical to the response mechanism, R_i, where $R_i = 1$ when unit i responds to the survey (and thus the outcome Y_i is observed for that unit), and 0 otherwise. That is to say that there is no nonresponse in a probability sample. Nonresponse occurs when $S_i = 1$ but $R_i = 0$.

The most basic estimator used to make population inferences about an outcome Y from probability samples is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952a). This estimator for a population mean $\bar{Y}_N$ is unbiased even when units are sampled with unequal probability[1]:

$$\hat{Y}_n = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i R_i}{\mathbb{P}(S_i = 1)} = \frac{1}{N}\sum_{\{R_i=1\}}\frac{Y_i}{\mathbb{P}(S_i = 1)} = \frac{1}{N}\sum_{\{R_i=1\}} Y_i d_i,$$

[1]Note that we follow the notation of Meng (2018) and include the population size N and sample size n when denoting population and sample quantities, respectively. For example, for a quantity Y, we will write the population and sample means as $\bar{Y}_N$ and $\bar{Y}_n$, respectively. As in Meng (2018), we do this to be explicit about the dependence on population size. We will explain how a core problem in survey research stems from a misunderstanding of this dependence.

where $d_i = 1/\mathbb{P}(S_i = 1)$ is the *design-based weight* for unit i, and is equal to the reciprocal of the design probability of selection.

For SRS, $\mathbb{P}(S_i = 1) = n/N$ for all units, so this estimator simplifies to the sample mean:

$$\hat{Y}_n = \frac{1}{N} \sum_{R_i=1} \frac{Y_i}{\mathbb{P}(S_i = 1)} = \frac{1}{N} \sum_{R_i=1} Y_i \frac{N}{n} = \bar{Y}_n$$

Horvitz and Thompson (1952a) also derive simple formulas for the variance of these estimators, so uncertainty from random sampling and the design-based weighting adjustments from unequal probability sampling can be easily quantified.

We can interpret each d_i as the number of population units represented by each sampled unit. Design-based weights d_i depend only on the sample design, and not the specific sample observed, since the weights are the reciprocal of the probability of selection, which must be known in advance.

In addition to SRS, there are a number of well-studied probability sampling designs, including stratified random sampling, cluster sampling, and systematic sampling. These designs are distinguished by how they assign selection probabilities $\mathbb{P}(S_i = 1)$ to units, and by the practical mechanisms used to select units. Sampling designs are generally chosen to suit particular applications in order to optimize the efficiency of the estimator (in terms of minimizing its variance), and to fit the practicalities of data collection in different settings.

The power of probability samples lies in the random selection mechanism. When units are selected at random from the population with (known and non-zero) probability, the resulting sample will be "representative" of the population of interest. By this we mean that the distribution of the set of *all possible* characteristics of a population, which we will denote $\mathbf{Z}$, is on average the same in the sample as in the population, and as a result, population inferences based on that sample are unbiased. Estimates from particular samples will vary from true population characteristics due to normal sampling variation, however this uncertainty is easily quantified using design-based estimates of estimator variance, like those of Horvitz and Thompson (1952a).

When this random selection mechanism is lost, so is any theoretical guarantee of the unbiasedness of sample-based estimators. This occurs when, for example, a sample suffers from nonresponse, and thus $S_i \neq R_i$. Nonresponse is controlled by individual respondent behavior rather than the researcher, so while $\mathbb{P}(S_i = 1)$ may be known, the probability of a unit responding $\mathbb{P}(R_i = 1)$ and observing Y_i for that unit is not know in advance. Researchers may check that respondents are "representative" of certain observed population characteristics and adjust for imbalances, but are limited by the characteristics that are observed in the sample and population, usually some small subset of $\mathbf{Z}$.

1.2.2 Selection bias

There are many types of non-sampling error that can cause surveys with probability sampling mechanisms to go awry. For example, measurement error occurs when the process of measuring an outcome changes how the outcome is observed and specification error may happen when a question does not actually measure the quantity it was intended to (Lohr, 2010). However, the error that I focus on here is what I believe to be the most pernicious type of error in modern survey research: that of selection bias. Selection bias occurs when the probability of *observing* a unit, $\mathbb{P}(R_i = 1)$, is 0 or not known in advance, for example if $R_i \neq S_i$.

Coverage bias is a type of selection bias that occurs when $\mathbb{P}(S_i = 1) = 0$, and thus $\mathbb{P}(R_i = 1) = 0$ as well (Lohr, 2010). In practice, this happens when, for example, the list from which a sample is drawn is an incomplete list of the population of interest. Consider a survey intended to measure public opinion among American adults that is conducted by soliciting respondents on Facebook. The list of potential respondents is limited to Facebook users, and thus American adults who are not Facebook users have $\mathbb{P}(S_i = 1) = 0$.

Nonresponse bias is another type of selection bias that occurs when units that are selected for a sample fail to respond, thus $S_i \neq R_i$. While the mechanism governing S_i may satisfy the requirements of a probability sample ($\mathbb{P}(S_i = 1) > 0 \; \forall i$ and $\mathbb{P}(S_i = 1)$ known in advance), the mechanism governing whether a unit is actually observed, R_i, does not as it is driven by individual behavior (Bradley et al., 2021). Thanks to public polling failures in 2016 and 2020, nonresponse bias is now more widely discussed outside of academic settings.

1.2.3 Types of Missingness

Broadly, selection bias refers to the bias that results from missing data. The causal inference literature defines three main types of missingness by the relationships between the response mechanism R, a specific outcome of interest Y, and a set of auxiliary covariates $\mathbf{X}$. The set $\mathbf{X}$ is a subset of all population characteristics $\mathbf{Z}$.

There are two parallel frameworks for understanding missing data and the resulting selection bias – the Neyman-Rubin potential outcomes framework (Rubin, 1974), and Pearl's Structural Causal Models (SCM) approach and the associated "do-calculus" conditions for recovery from selection bias (Pearl, 1995a). Rubin's framework tends to dominate in the social sciences, while Pearl's framing is more common in computer science. I refer you to Chapter 5 for a more comprehensive overview of Pearl. Here I will rely primarily on Rubin, but will mention the

analogous do-calculus condition. The three types of missing data are 1) missing completely at random (MCAR) 2) missing at random (MAR) and 3) missing not at random (MNAR).

Missing completely at random (MCAR), $R \perp\!\!\!\perp Y$: If R is completely independent of Y, $\mathbf{Z}$, and the sampling design, then units are MCAR. In this case, we can think of units that were excluded from the sample or that failed to respond as an SRS from the original sample. The respondents will, on average, be representative of the population of interest. If the nonresponse mechanism is unknown (and therefore the sample is not a probability sample), but if it is completely orthogonal to both $\mathbf{X}$ and outcomes of interest, then population estimates will not be biased.

Missing at random (MAR), $R \perp\!\!\!\perp Y|\mathbf{X}$: If the response mechanism depends only on observed auxiliary covariates $\mathbf{X}$, then units are MAR. As long as we adjust for the $\mathbf{X}$ in analysis, we can still derive unbiased estimates of Y using various adjustment methods (discussed in the following section), making nonresponse *ignorable*.

Missing not at random (MNAR): If the researcher cannot identify a set $\mathbf{X} \subseteq \mathbf{Z}$ such that $R \perp\!\!\!\perp Y|\mathbf{X}$, or cannot observe elements of $\mathbf{X}$ in one of the sample or the population, then units are MNAR. The inability to identify a sufficient $\mathbf{X}$ could be the result of direct dependence between R and Y that is not mediated by other mechanisms, which would be impossible to adjust for, even if every element of $\mathbf{Z}$ was perfectly observed. This is the most problematic type of missingness.

As Lohr (2010) says, "The best way to deal with nonresponse is to prevent it." However, despite researchers' best efforts, it is almost impossible to prevent nonresponse, and selection bias more broadly, in practice. Instead, we generally rely on methods, like **weighting**, that assume a model for the response mechanism to adjust for selection bias after it has occurred.

The main difficulty of selection bias in practice, in both probability and nonprobability samples, is that researchers never know for certain what type of missingness is occurring, and it is only possible to recover from selection bias in settings where data is MCAR or MAR. Here we mean "recover from selection bias" in the sense of Pearl (1995a); that the response mechanism and observed data meet a set of conditions such that we can still derive unbiased population inferences from the data.

Furthermore, the type and degree of missingness depends on the outcome of interest and chosen estimator. A set $\mathbf{X}$ that is sufficient to ensure $R \perp\!\!\!\perp Y|\mathbf{X}$ may not be sufficient for a different outcome. When there is no selection bias, the properties of probability samples ensure that *all* measured outcomes are uncorrelated with the response mechanism, and thus all estimators of populations quantities of those outcomes will be unbiased.

1.2.4 Nonprobability samples

Nonprobability samples are the set of samples that do not meet the definition of a probability sample, and therefore suffer from selection bias. There are myriad types of samples that do not fit the definition of a probability sample, however, colloquially, "nonprobability samples" often refer to **online opt-in surveys**. These surveys rely on a panel of participants that have previously indicated willingness to participate in surveys. While these panels may be large, they still rarely cover the entire population of interest. Take YouGov, a well-respected online panel vendor, for example. As of 2021 (more recent data is not publicly available), YouGov's US panel included 2 million potential respondents, which is still less than 1% of the US adult population (YouGov, 2021). American adults who have not been recruited into YouGov's panel have no chance of being selected to participate in one of their surveys, $\mathbb{P}(S_i = 1) = 0$, which violates the definition of a probability sample.

Other nonprobability designs commonly found in survey research include **quota sampling**, in which a set of population targets, or quotas, are defined using dimensions that the researcher believes are sufficient to achieve conditional independence of an outcome of interest and the response mechanism (Lohr, 2010). Respondents are sorted into strata based on $\mathbf{X}$, and invited to participate using some non-random mechanism until the joint distribution of $\mathbf{X}$ in the sample matches that of the population. Inferences will be unbiased if $\mathbf{X}$ is truly sufficient to ensure conditional independence of the response mechanism and the outcome of interest.

Administrative data is a particular type of nonprobability sample that is becoming more readily available to researchers, but one in which the impact of a biased response mechanism is often underappreciated. A researcher would likely immediately and intuitively see the flaws in using a 5% of the population nonprobability sample (without adjustment) to make population inferences. However, it may be less intuitive that a nonprobability sample that makes up 95% of the population may be as unsuitable for making population inferences as the 5% sample.

Meng (2018) was motivated exactly by the question of whether data quantity can make up for the lack of a probabilistic response mechanism, and proves mathematically that it can be highly inefficient to try to overcome the lack of a probabilistic response mechanism by collecting (even a lot) more data. We will discuss this framework in more detail in Section 1.2.6.

Large health studies, like the UK Biobank, a prospective health study in the UK with over 500,000 participants (Sudlow et al., 2015), are another type of nonprobability sample in which selection bias is often overlooked. Often these studies seek to estimate associations between health characteristics rather than

point estimates in the population, so researchers assume that selection bias is not important to consider (Fry et al., 2017). As we will show in Chapter 5, this assumption is incorrect. Benonisdottir and Kong (2023) even find that genetic traits can actually impact the likelihood of participation in genetic studies, a clear opportunity for selection bias to impact estimates of health outcomes.

According to our definition of a nonprobability sample, even a sample that was drawn probabilistically but suffers from nonresponse (even of a single unit) is actually a nonprobability sample because the response mechanism is not purely random, but rather is impacted by individual response behavior. However, the robustness of probability sampling theory to selection bias depends on its type and degree.

1.2.5 Adjusting for selection bias

Completely preventing selection bias requires perfect sampling frame coverage and the observation of every selected unit, which is nearly impossible in practical settings. Instead, in order to leverage the nice theoretical properties of probability samples, missingness must be MCAR or at least MAR with a sufficient auxiliary set. Generally, researchers assume the latter and try to select an appropriate $\mathbf{X}$. In order to successfully recover from selection bias under MAR, researchers must:

1. Develop a model for the response mechanism R
2. Identify a sufficient auxiliary set $\mathbf{X}$ such that $R \perp\!\!\!\perp Y|\mathbf{X}$
3. Observe $\mathbf{X}$ in both the survey and the population
4. Analyze Y in a way that adjusts for $\mathbf{X}$.

However, this is more easily said than done. The following section describes each of the steps above, and their associated challenges, in greater detail.

Modeling the response mechanism R

Correctly modeling R requires extensive prior knowledge about the survey design, response patterns, the specific research context, and luck. Response mechanisms are context-specific, depend on the survey mode, and even the researcher conducting a study (Groves and Peytcheva, 2008). Response mechanisms are also highly heterogeneous across individuals and can vary within a single survey depending on the outcome of interest (e.g. item nonresponse that results from social desirability or poorly worded questions) (Groves and Peytcheva, 2008). Perhaps most

problematically, response mechanisms are *dynamic* over time, even controlling for all the factors just mentioned. See the dynamic partisan nonresponse bias in Gelman et al. (2016) for an example.

Pearl (1995a) introduces Structural Causal Models (SCMs) as powerful tools for encoding complex dependence structures and prior contextual knowledge and specifically demonstrates their utility for capturing complex relationships between response mechanisms, outcomes, and other relevant mechanisms. For a full introduction to the extensive literature on SCMs, see Pearl et al. (2016), Spirtes et al. (2001), and Morgan and Winship (2007), or a primer in Chapter 5. Pearl's work focuses primarily on causal outcomes, while most of the surveys we consider in this work do not seek to make causal inferences. However, Bareinboim et al. (2014) derives SCM-based conditions for observational outcomes that are less strict than those required for causal inferences.

In large-scale public opinion research, to my knowledge, using SCMs is not standard practice. While pollsters do not, to my knowledge, use SCMs explicitly, the process of selecting features to use in adjustment often involves an informal, *implicit* model of the response mechanism. For example, in the 2020 presidential election, some states allowed voters to vote prior to election day, either in person or by mail. Some states reported the number of people who had "early voted" prior to election day, and others even reported who had already voted at the individual-level using publicly available voter files. It became clear that the people who were voting early were both more likely to respond to surveys and more likely to be stronger partisan Democrats than Democrats who did not vote early. This resulted in large selection bias in many pre-election polls, including in one Washington Post poll from late October which predicted that Joe Biden would win Wisconsin by 17 percentage points (Clement et al., 2020), when less than a week later, Biden went on to win Wisconsin by a margin of only 0.6 percentage points (Wasserman et al., 2020). Some pollsters, like the New York Times / Sienna College (Cohn, 2020), noticed prior to election day the concerning correlation between response and strength of Biden support among "early voters" and, even more troublingly, that it persisted after weighting on their usual auxiliary set. Therefore, they adjusted their model of response in-cycle and began to weight on "early vote status."

Whether one is using SCMs or another framework for encoding the model for R, the most common, and perhaps most dire, error in this step is omitting key mechanisms driving R. Methods, like Egami and Hartman (2021) who propose a data-driven algorithm for modeling R, are only able to consider features that have already been observed in the sample and the population. Furthermore, empirical

methods for learning a model for R require the researcher to have some understanding of the units that do not respond to a survey, which, by definition, is difficult.

A crucial example of this type of error was the failure of many pollsters to recognize that education was a key driver of nonresponse in 2016 election polls (Kennedy et al., 2018). In 2016, white Americans without college degrees were both less likely to respond to surveys, and more likely to support Donald Trump. This nonresponse pattern, and relationship between education and partisanship, had not been prevalent in previous US political polling, so pollsters were largely unaware of its relevance. As a result, many pollsters specified an incorrect model for R that did not depend on education, so did not use it in adjustment and thus missingness was MNAR rather that MAR. This resulted in widespread shock when Donald Trump won the presidential election despite most forecasts based on pre-election polling had predicted 70% to as high as 99% chance of a Clinton win (Kennedy et al., 2018).

Identifying the auxiliary set X

If the researcher does in fact manage to perfectly specify a model for the response mechanism, the next obstacle lies in correctly defining $\mathbf{X} \subseteq \mathbf{Z}$, to ensure that $\mathbf{X}$ is sufficient to ensure conditional independence of R and Y. Pearl (1995a) uses the language of SCMs to describe $\mathbf{X}$ as the set of mechanisms that *d-separates* the selection mechanism R from the outcome of interest Y. This means that $\mathbf{X}$ is the set of variables that blocks all paths in the SCM between R and Y such that no perturbation of one has an impact on the other.

Unfortunately, this is not as simple as identifying every feature that contributes to the response mechanism and the outcome, because this may lead to another form of bias, called *collider bias*. Collider bias occurs when we condition on the mutual effect of two variables or its child, thus inducing spurious correlation between the two mutual effects. Nguyen et al. (2019) demonstrates with simulations that the direction and magnitude of the bias depend on the true relationship between mutual effects and other factors, while Munafò et al. (2018) and Day et al. (2016) give examples of problematic collider bias in practice. In the context of political pre-election polling, collider bias might occur if, for example, political affiliation is the mutual effect of educational attainment and an individual's opinion about abortion access. If a survey measuring those three features is weighted using (and thus conditioned on) education and political affiliation (as is common practice), then collider bias could impact estimates of the association between education and attitudes toward abortion access. It is not enough to identify all the features in $\mathbf{Z}$

that are correlated with R and Y, the researcher must also take care to explicitly exclude from $\mathbf{X}$ any potential colliders.

Furthermore, verifying that the auxiliary set has been correctly specified is nearly impossible. One intuitive approach to verifying that $\mathbf{X}$ meets the conditional independence criterion, $R \perp\!\!\!\perp Y|\mathbf{X}$, is to test for the residual dependence between R and Y. However, there are two problems with this approach. First, to test $R \perp\!\!\!\perp Y$, we must observe some Y for units where $R_i = 0$, which by definition is not possible for the sample of interest. Second, Meng (2018) gives striking examples of how even tiny correlation (on the order of 0.005) between outcome and response mechanism can have a catastrophic impact on estimator bias. Testing for independence would require a test sensitive enough to consistently detect correlation of this magnitude.

Identification of $\mathbf{X}$ is therefore typically a highly manual endeavor based on the prior knowledge and domain expertise of the researcher. There are some methods that treat selection of $\mathbf{X}$ as a variable selection task, namely Caughey and Hartman (2017) and my work Chapter 5 that modifies Caughey and Hartman (2017) to improve stability of selection of $\mathbf{X}$ and resulting estimates. However, employing these methods without considering the causal dependence structure of the response mechanism runs the risk of accidentally inducing collider bias. Some manual checking remains necessary.

Observing the auxiliary set X

One final practical limitation is the requirement of observing the distribution of $\mathbf{X}$ in both the sample and the population. Observing auxiliary variables in the sample is straightforward - researchers can add questions to a survey intended to measure quantities that they believe may be important auxiliary variables (e.g. age, gender, education, race, etc.). As long as questions are designed appropriately and $\mathbf{X}$ identified correctly in advance of data collection, these features should be reliable. It is important to note that this step is not always as straightforward as it may seem. See Kennedy (2020) for an example of the complexities of collecting information about sex and gender, and similarly race/ethnicity, in surveys.

Perhaps a larger challenge is observing the necessary features in the population. Generally survey researchers rely on government censuses or commercially-acquired individual-level consumer data files to define the population. In each case, $\mathbf{X}$ must be observed for the entire population or imputed (which presents its own challenges and potential for error). The response scale used to measure features in the auxiliary set must be the same as those used in the sample. For example,

the categories offered to respondents when asked about their race in a survey must match those reported in the census.

Additionally, population data is measured infrequently – censuses only occur once per decade in the UK and the US – so high-quality population data may be quite outdated. It also may only be available in aggregate (e.g. in the form of pre-specified tables), and population information about specific key joint distributions of elements in $\mathbf{X}$ may not be available to the researcher. If individual-level population data is available, generally only a small, properly-anonymized sample is available to the public. If a researcher requires more up-to-date population data than the most recent census, an alternative is to use a high-quality large-scale survey (like the American Community Survey in the US or the Annual Population Survey in the UK). However, one must then account for the uncertainty in population benchmarks as well as those in the survey of interest in population inferences. Population frame uncertainty is rarely discussed, let alone quantified, in survey research. I seek to address this in this thesis – Chapter 2 introduces the concept of "benchmark uncertainty intervals" and Chapter 4 proposes a method for quantifying the sensitivity of weighted estimators to uncertainty in population targets.

One tactic that well-resourced researchers and commercial firms have used to alleviate some of these challenges is to require that survey responses can be matched back to a database that includes individual-level features for the entire population of interest, as described in detail by Ghitza and Gelman (2020). This ensures that the features used for adjustment are measured identically between the survey and population, and opens up a wealth of potential features (e.g. political affiliation) for which public data may not be readily accessible. The downside, however, is that it limits the modes that can be used for survey research because respondents must be recruited in such a way that they can be linked back to the database, or surveys must collect invasive personal information to be used in a complex matching process in order to probabilistically match respondents back to individuals in the database.

Another approach researchers have taken involves synthesizing population data from a variety of sources, using many of the same methods developed to adjust samples for nonresponse. The approach is heavily utilized by YouGov and described in more detail in Lauderdale et al. (2020).

Methods for adjustment

With all the aforementioned limitations and challenges in mind, we now arrive at the final step in recovering from nonresponse bias – adjusting survey responses using the identified auxiliary set $\mathbf{X}$. There are three broad categories of algorithms used for adjustment: 1) Inverse probability weighting (IPW), 2) directly modeling outcomes of interest, and 3) matching.

Inverse Probability Weighting (IPW)

This technique stems directly from the Horvitz-Thompson estimator (Horvitz and Thompson, 1952a) which uses the inverse of the probability of selection to derive unbiased estimators for unequal probability samples. IPW weights, w_i, modify the design weights, $d_i = 1/\mathbb{P}(S_i = 1)$, to account for nonresponse. IPW assumes that although we may not know the probability of response in advance, we can estimate it empirically with the auxiliary set. IPW estimators for population means stem directly from Horvitz-Thompson estimators, but use the derived weights w_i instead of the design weights d_i (Deville et al., 1993):

$$\hat{Y}_N = \bar{Y}_w = \frac{\sum_i Y_i R_i w_i}{\sum_i R_i w_i} \tag{1.1}$$

Deville et al. (1993) derive a class of such algorithms that minimize the distance between $\mathbf{w} = \{w_i, \ldots, w_N\}$ and $\mathbf{d} = \{d_1, \ldots, d_N\}$, defined by some distance measure D, subject to the constraint $\sum_N R_i w_i \mathbf{X}_i = \sum_N \mathbf{X}_i$. Specific weighting algorithms are defined by the choice of distance measure D and the method of minimization. This class of IPW algorithms includes raking, post-stratification, and calibration, three of the most common IPW methods used in practice.

IPW weights are flexible and can be easily incorporated into estimators of population means and totals and regression models (Lohr, 2010, Chapter 11). As long as selection bias affects outcomes in a survey similarly, a single set of weights is sufficient for analysis of all outcomes measured in that survey, and their joint distributions. Hernán et al. (2004) describes this property by saying that IPW "creates a pseudopopulation" that does not suffer from selection bias.

The classical IPW methods of Deville et al. (1993) suffer from some key limitations. First, there is a limit to the size of the set $\mathbf{X}$ that most algorithms can accommodate, either due to computational constraints (raking, calibration), or due to empty strata, as defined by the interaction of all elements in $\mathbf{X}$, in the sample (poststratification). Second, raking and poststratification can only accommodate categorical $\mathbf{X}$, so continuous variables must be discretized. Calibration, the only method that can accommodate continuous variables, can be highly unstable.

One could also seek to estimate the probability of response $\mathbb{P}(R_i = 1)$, which we will denote π_i, directly with any sort of predictive model, then estimate the IPW weights as $w_i = 1/\hat{\pi}_i$. The drawback to this approach is that there is no guarantee that the sample totals of auxiliary variables will match those in the population, and most standard modeling techniques would require rich information on unobserved units.

Plenty of research is aimed at addressing those limitations. Caughey and Hartman (2017) introduce raking on all 2-way interactions of features in $\mathbf{X}$ as a way to leverage the strengths of both raking and poststratification. Though not designed as a method specific to survey research, Kernel Mean Matching (KMM) from Gretton et al. (2013) presents a nonparametric framework for IPW, noting that traditional post-stratification can be described as KMM with a linear kernel. Kern et al. (2020) examines the utility of methods that combine kernel weighting with various machine-learning methods, like gradient tree boosting and conditional random forests, for adjusting for nonresponse in probability samples. I explore a related approach in Chapter 5.

When using IPW methods, one must account for the inherent uncertainty in the weighting procedure when making population inferences. Closed-form variance formulas are available for the Horwitz-Thompson estimator, and some IPW weighting methods (Lu and Gelman, 2003), though these formulas are less user-friendly. In practice, researchers often rely instead on bootstrapping, replicate weights (Fay and Train, 1995), or design effects (Kish, 1992) to capture the additional variance from weighting.

These methods account for the increase in variance due to adjustment needed when sample distributions of $\mathbf{X}$ differ from those of the population, but *do not account for the cause of that discrepancy*. Consider a simple example. Say we are interested in estimating $\bar{Y}_N$ in some population, and observe two samples with sampling mechanisms S^1 and S^2, and response mechanisms R^1 and R^2. Both samples are of size n, such that $\sum_i R_i^1 = \sum_i R_i^2 = n$, and we observe Y_i and X_i for units where $R_i = 1$ in each sample.

The response mechanism in the first sample, R_i^1 is probabilistic, but sampled with unequal probability defined by $\mathbf{X}$ such that $\mathbb{P}(S_i^1 = 1) = \mathbb{P}(R_i^1 = 1) = f(\mathbf{X})$. Thus, $\mathbb{P}(X|R_i^1 = 1) \neq \mathbb{P}(X)$, and unbiased population inference requires adjustment with respect to $\mathbf{X}$. In the second sample, the sampling mechanism S_i^2 is probabilistic, but the sample suffers from nonresponse, and thus $\mathbb{P}(S_i^2 = 1) \neq \mathbb{P}(R_i^2 = 1)$, and $\mathbb{P}(R_i^2 = 1)$ is some unknown function of $\mathbf{X}$ and unobserved features $\mathbf{U}$.

By chance, the two samples have the same distribution of $\mathbf{X}$, such that $\mathbb{P}(X|R_i^1 = 1) = \mathbb{P}(X|R_i^2 = 1)$, and the researcher applies identical IPW adjustment procedures. While the uncertainty in the first sample can be quantified using the well-studied mathematical properties of probability sampling and Kish's design effect from weighting, the uncertainty in the second sample depends on an unknown mechanism and unobserved features $\mathbf{U}$, so cannot be quantified. We *assume* that X is sufficient to induce conditional independence, but we do not know for certain.

Directly modeling outcomes

Another category of adjustment methods involve modeling an outcome of interest directly, incorporating features from the auxiliary set. Perhaps the most well-known of these methods is multilevel regression and poststratification (MRP) (Gelman and Little, 1997; Park et al., 2004). MRP first estimates a predictive model for a particular outcome, then projects that outcome onto a representative population frame. The predictive model is often a multi-level model due to its powerful partial pooling properties and ability to capture the hierarchical structure (e.g. geographic) that exists in most applications of MRP. The method is particularly powerful for deriving small-area estimates (e.g. US states or UK constituencies) when survey data is only available at the national level.

The literature on MRP, its applications, and extensions is vast. Though multi-level models are most common, they are by no means necessary. Similarly, any representative population frame in which the auxiliary variables of $\mathbf{X}$ are observed can be used at the projection step. For example, Ghitza and Gelman (2020) uses an individual-level population frame. The poststratification step in classical MRP requires that model and population features are categorical, but Gao et al. (2021) introduces structured priors that can capture the correlation structure between levels of ordered factor variables or buckets of discretized continuous predictors.

While this approach supports more complex estimation techniques for Y, it must be applied to outcomes separately. This is far too computationally intensive for most traditional public opinion pollsters, whose polls can have upwards of 50 questions and need to release estimates quickly. Some, however, do employ more automated methods that can scale with questionnaire length more easily.

MRP and other model-based methods of analysis still rely on assumptions that the models for Y and R have been correctly specified, and the adjustment set $\mathbf{X}$ correctly identified.

Matching

The last method, described in detail in Schaffner et al. (2019), involves matching an unrepresentative sample of survey respondents to a true probability sample drawn from the target population. Each unit in the target sample is matched to a respondent from the nonprobability sample. Matching can be done exactly, using propensity scores (modeled probability of being observed), or proximity matching (minimizing some distance metric), using the set of auxiliary features $\mathbf{X}$. Respondents that are not matched to the target sample are discarded.

This method improves upon IPW because analysis of the matched sample is not limited by the need to incorporate weights into estimators - the matched sample is itself representative. Matching also improves upon modeling the outcome directly in that it only needs to be performed once.

In order to implement matching, the researcher needs to observe individual-level population data to match survey responses to. This is a limitation for less well-resourced researchers. Another drawback of matching is that one is forced to discard survey responses for which there is no suitable population match, for example if the sample contains too many similar respondents who are over-represented relative to the population. Discarding sample data can feel wasteful, especially when responses are costly to acquire, and doing so decreases the size of the sample used for analysis, increasing estimator variance. However, this bias-variance trade-off is seen in any adjustment method.

1.2.6 Measuring selection bias with *ddc*

Thus far we have discussed probability samples that form the foundation of survey statistics, nonprobability samples that suffer from selection bias, and the forms that missingness can take. We have also introduced a series of methods for recovering from observed selection bias, assuming that missingness is MAR. However it is difficult, if not impossible, in most practical settings to know for sure whether the assumption of MAR holds, and even if it does, the degree to which adjustment was successful. Meng (2018) establishes a mathematical framework that allows us to examine some of these questions.

Meng (2018) derives the following identity relating error in the sample estimator $\bar{Y}_n$ or a population mean $\bar{Y}_N$ to three intuitive quantities:

$$\bar{Y}_n - \bar{Y}_N = \rho_{Y,R} \times \sqrt{\frac{1-f}{f}} \times \sigma_Y \tag{1.2}$$

where $f = n/N$ is the sampling rate and σ_Y is the population standard deviation of the outcome of interest Y. The final term, $\rho_{Y,R} = \text{Corr}_J(Y_J, R_J)$, is the correlation

between the outcome and response mechanism according to the uniform distribution on random index J defined on $\{1, \ldots, N\}$. The power of this identity lies in the simplicity and interpretability of its components. Meng offers the following descriptions for the three quantities:

- *Data quantity*: the contribution of data quantity to estimation error is measured by the dropout odds $\sqrt{\frac{1-f}{f}}$. This quantity will be exactly 0 when the entire population is observed ($f = 1$), and infinite when no data is observed.

- *Problem difficulty*: σ_Y, the standard deviation of Y_J, which captures the impact on estimation error of innate problem difficulty. If $\sigma_Y = 0$ and Y is constant in the population, then error goes to 0. When σ_Y increases, so does estimation error.

- *Data quality*: Meng (2018) introduces a new quantity, $\rho_{Y,R}$, or the **data defect correlation**, which measures the sign and degree of selection bias caused by the response mechanism R. Recall that when there is no nonresponse in a probability sample, or if nonresponse is MCAR, then an estimator Y is completely independent of the response mechanism R, implying a correlation of 0, and therefore 0 estimation error. As correlation increases, so does estimation error.

Meng (2018) uses the identity in Equation 1.2 to derive the following decomposition of MSE for the estimator $\bar{Y}_n$:

$$\text{MSE}_R(\bar{Y}_n) = \mathbb{E}[\bar{Y}_n - \bar{Y}_N]^2 = \mathbb{E}[\rho_{Y,R}^2] \times \frac{1-f}{f} \times \sigma_Y^2. \tag{1.3}$$

The key difference here is the expectation of $\rho_{Y,R}$ which describes the expected behavior of a response mechanism R rather than a particular realization of it. The quantity $\mathbb{E}[\rho_{Y,R}^2]$ is called the data defect index, or *ddi*. While the *ddc* is defined from -1 to 1 and indicates sign and direction of bias in a particular sample, the *ddi* ranges from 0 to 1, with larger values indicative of lower data quality in general for response mechanism R and estimator $\bar{Y}_n$.

This identity can be extended to account for the impact of IPW used to adjust for selection bias:

$$\bar{Y}_w - \bar{Y}_N = \rho_{Y,R_w} \times \sqrt{\frac{1-f_w}{f_w}} \times \sigma_Y \tag{1.4}$$

where Y_w is the weighted sample mean, $f_w = n_w/N$ where n_w is the effective sample size that accounts for variance inflation from weighting using Kish's design effect, $n_w = n/\text{deff}$ where $\text{deff} = 1 + \frac{\bar{w^2}}{\bar{w}^2}$, and R_w is the weighted response indicator, $R_w = R_i w_i$. This modification captures the bias-variance trade-off of using IPW to lessen the impact of selection bias – weighting aims to make use of the MAR assumption, that $Y \perp\!\!\!\perp R|\mathbf{X}$, so constructs the weights using $\mathbf{X}$ which ensures $Y \perp\!\!\!\perp R_w$ and hence decreases $\text{abs}(\rho_{Y,R_w})$. However this comes at the cost of inflated variance, expressed through the decrease in effective sample size and sampling fraction.

There are a number of key implications of these identities, which Meng (2018) elegantly examines in detail, but which I will, less elegantly, summarize here:

- Holding sample size constant, estimator error depends in part on population size. When we rewrite the data quantity term as $\sqrt{(N-n)/n}$, it becomes more clear that fixing n, estimator error will increase with N. Therefore, quantity of data should be measured relative to population size, with sampling fraction n/N, instead of absolute size n.

- Probability samples control estimator error by controlling $\rho_{Y,R}$ at a rate of $N^{-1/2}$, or equivalently $\mathbb{E}[\rho^2_{Y,R}]$ at a rate of N^{-1}, thus guaranteeing high data-quality and eliminating the dependence of estimator error on population size.

- *Law of Large Populations*: For studies with the same expected data defect correlation $\mathbb{E}_R(\rho_{Y,R}) \neq 0$, the stochastic error of $\bar{Y}_n$, relative to its benchmark under SRS, grows with the population size at rate $N^{-1/2}$.

- The bias-adjusted effective sample size of a nonprobability sample (the size of a probability sample that would be expected to produce the same level of error) is inversely proportional to $\mathbb{E}[\rho^2_{Y,R}]$. More precisely, it is $n_{\text{eff}} \leq \frac{f}{1-f}\mathbb{E}[\rho^2_{Y,R}]^{-1}$

- The data quality index depends on a specific estimator, and can vary even within a sample for different outcomes. This is similar (and fundamentally related to) the concept that type of missingness is dependent on a specific outcome.

While these observations may seem simple, Meng (2018) shows that the *scale* of the problem that nonprobability samples face in controlling estimation error is striking. In practical terms, say we want to estimate support nationally for Joe Biden in advance of the 2020 US presidential election. The population of eligible voters in the US is about 240 million. In order for a nonprobability sample to have

the same level of error as a similarly-sized probability sample, we would need to ensure that $\rho_{Y,R} \leq 6.5 \times 10^{-5}$. Alternatively, say that we observe a 1% sample of the population for which $\rho_{Y,R} = 0.05$. In that case, the effective sample size of this data is $n_{\text{eff}} \leq \frac{0.01}{0.99} \times 0.05^{-2} = 4$ (not a typo!). If the sample was instead 50% of the population, the effective sample size would only increase to $n_{\text{eff}} \leq \frac{0.5}{0.5} \times 0.05^{-2} = 400$. These results may seem extreme, but as Meng (2018) notes,

> It is indeed extreme, but what should be unbelievable is the magical power of probabilistic sampling, which we all have taken for granted for too long.

This framework is powerful and, to date, underutilized. One of the main advantages of this framework is that the *ddi* $\mathbb{E}[\rho^2_{Y,R}]$ can be compared across studies and outcomes. The main challenge of this framework is that in order to estimate $\mathbb{E}[\rho^2_{Y,R}]$, you must observe the population benchmark $\bar{Y}_N$, which is rarely possible in practice. Even in the election example given above, we only observe the population truth after the election is over, when we are no longer as interested in estimating Biden's projected margin. In that case, we are able to estimate *ddi*, but are unable to use it as a tool to correct for selection bias proactively. Isakov and Kuriwaki (2020) employ one strategy for estimating $\mathbb{E}[\rho^2_{Y,R}]$ without observing $\bar{Y}_N$ by using a Bayesian hierarchical model to learn the structure of $\mathbb{E}[\rho^2_{Y,R}]$ in a similar setting where benchmark data is observed, then apply that structure to the survey data of interest. However, this approach requires the assumption that the model of response is the same in both settings.

We describe in Bradley et al. (2021) (Chapter 2) how *ddc* can also be decomposed by stages of the data collection process, e.g. mode selection, sampling frame definition, sample design, fieldwork, and analysis. If estimates of true error are available for each stage of the data collection process, then it is possible to estimate the contribution that each stage makes to the net selection bias observed in the final sample. Even if true error is not observed for each stage of data collection, the *ddc* decomposition by stage still provides valuable insight into the stages at which non-random response mechanisms are most destructive to overall sample quality.

We show that, just as the overall population size can magnify the impact of *ddc*, the impact of the *ddc* at each stage of data collection is moderated by the relevant population size at that stage. For example, the impact of nonresponse bias (as measured by *ddc*) at the fieldwork stage depends on the size of the sample rather than the overall population size. As the relevant population size decreases at each stage of the data collection process, the stage that dominates overall *ddc* is the first stage at which the random selection mechanism is lost. In practice, this implies

that selection bias introduced in the mode selection and sampling frame definition stages is not then mitigated by a random selection mechanism for selecting potential respondents in a subsequent stage of data collection.

1.2.7 From "probability vs nonprobability" to "total selection bias minimization"

So far we have given a brief overview of the core theoretical foundation of modern survey research. In this next section, I hope to convey the monumental challenge that researchers face in trying to produce unbiased population estimates based on survey data. These obstacles are not insurmountable, but consistently producing high-quality survey data requires more than simply deciding that "probability" or nonprobability samples with complex analysis is the preferred approach. Instead, it requires a relentless and comprehensive approach to addressing selection bias. We will mainly examine surveys in the context of opinion polling in the US, and note that the specifics may vary by country and research context.

Figure 1.1 outlines the points at which selection bias can enter the data collection process, and outlines the key methods at each stage for mitigating it. This framework builds heavily on the Total Survey Error framework (Biemer and Lyberg, 2003), but focuses exclusively on selection bias rather than the other sources of survey error. The rest of this section describes this framework in greater detail.

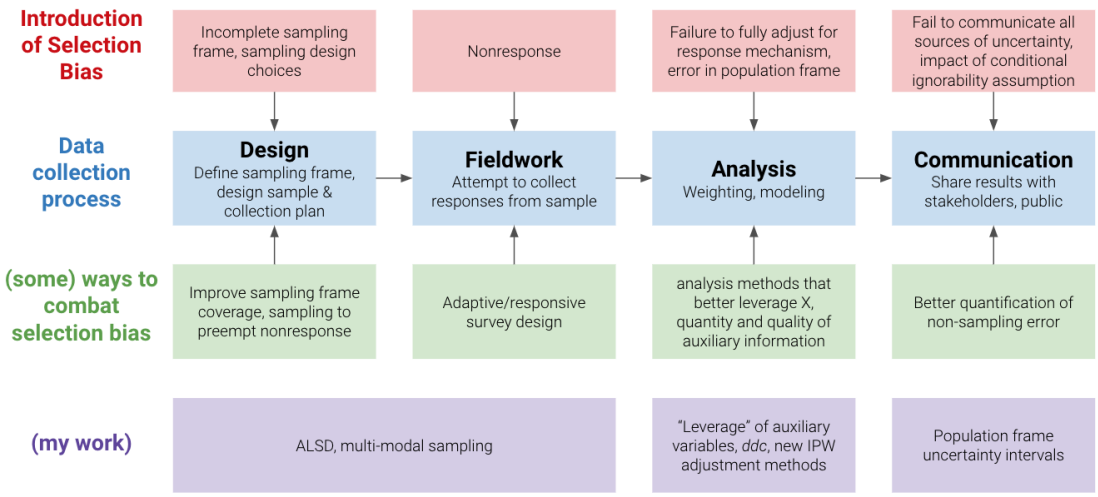


Figure 1.1: Framework for understanding where and how selection bias is introduced into observed samples, and where my research contributes.

Design

There are a number of ways in which selection bias can enter a sample in the design stage. First, selecting the mode used to collect responses can systematically exclude people who are not contactable at all. For example, a cell phone-only survey will by definition exclude the approximately 1% of American adults that do not have a cell phone (Blumberg and Luke, 2022). Thus, even random digit dialing of landline and cell phone numbers, which does not require a list of the population in advance, would give over 2.8 million US adults a probability of selection equal to 0.

Similarly, an online survey necessarily excludes people who lack the ability to access the internet. The American Community Survey conducted annually by the US Census Bureau estimated that in 2016, 89% of households had a computer or a smartphone, and 81% of households had broadband internet Ryan (2017). Rates of computer and smartphone ownership varied with age, race, urbanicity, and income. Almost 20% of adults in the US had no access to the internet at home in 2016. While adults may have internet access in places outside their homes, the lack at home suggests a significant obstacle to participation in online surveys.

Once a mode (or combination of modes) has been selected, defining the sampling frame may unintentionally introduce selection bias. For phone surveys, lists of phone numbers must be acquired, generally from some commercial data vendor or voter registration database. These numbers are often sourced from credit agencies, so may be lower in quality for lower-income and more transient populations. Although defining a sampling frame is not an obvious part of running an online survey, selecting an online panel vendor implicitly defines the sampling frame as the list of their panelists.

Some online surveys, like Pew Research's American Trends Panel, ensure that individuals who do not have internet in their homes are included in the panel by providing recruited respondents with tablets and internet access Pew Research Center (2020). However, the opaqueness of the recruitment methods of many online panels makes it difficult to evaluate how comprehensive their sampling frames are. YouGov, one of the largest and most well-respected online survey firms, has only 2 million panelists in the US from which to draw samples, or about 1% of the US adult population (YouGov, 2021). It is fair to assume that other online panels are not significantly larger than YouGov's.

Of the modes most commonly used in survey research – landline telephones, cell phones, online surveys, SMS – not one has a comprehensive sampling frame. Face-to-face and mail surveys theoretically have complete sampling frames (the addresses of everyone in a particular area), but are rarely used in survey research in

the US due to the financial and logistical cost of implementation. Some large-scale government surveys, like the decennial US Census and the Current Population Survey (CPS), are exceptions and are entirely face-to-face or include face-to-face components, however few other institutions choose to conduct face-to-face surveys when telephone and internet surveys are far cheaper and faster.

If random sampling is used at all, it is only after mode and sampling frame design choices have been made. The sampling stage is the stage at which the researcher has the greatest control over selection bias, however selection bias may still be introduced inadvertently. For example, selecting a sample may include assigning sampled individuals to different modes of outreach. If this assignment is not probabilistic in nature, it may introduce systematic bias.

Fieldwork

The primary source of selection bias in the fieldwork stage is nonresponse. As discussed previously, response rates across survey modes have markedly declined over the past three decades. In 2018, Pew Research observed average telephone response rates of 6%, a decline from 15% in 2009 and 36% in 1997 (Kennedy and Hartig, 2019). In 2020 and 2021, the US Census Bureau conducted the Household Pulse Survey, intended to measure the impact of COVID-19 on Americans and observed survey response rates of only 4% to 8% (US Census Bureau, 2021). There are a number of hypotheses for why this may be the case:

1. The sharp increase in spam calls on both landlines and cell phones, making people less likely to answer calls from unknown numbers,
2. Polling failures in 2016 led to lower confidence in polling in general and less willingness to participate,
3. (In the US) Donald Trump's tendency to sow doubt in institutions among supporters,
4. Growing concerns about data privacy,
5. New technologies intended to prevent spam calls occasionally block pollsters by accident.

Not only is nonresponse getting worse overall, but it is dynamic. To again refer to the 2016 US presidential election, pollsters failed to weight on education because it had not been as critical prior to a political realignment between 2012 and 2016 along education lines. Furthermore, a meta-analysis by Groves and

Peytcheva (2008) shows significant within-study heterogeneity of nonresponse bias, indicating that individual-level drivers of nonresponse bias are often larger than study-level ones. This indicates that nonresponse is very unlikely to be missing completely at random within a particular study.

Analysis

It is possible to introduce additional selection bias in the analysis stage – either by systematically excluding certain units from analysis, or by accidentally introducing bias by conditioning on a collider – however the main challenge that it poses is an inability to properly eliminate it.

First, survey researchers must perfectly assess a model for heterogeneous nonresponse within a survey, on the individual level. Large polling misses, like the 2016 US presidential election, should convince you that this is hard to do in practice. Over half of the polls fielded in the last two weeks of the 2016 election failed to foresee the importance of education to the model of response in 2016 and account for it in weighting (Kennedy et al., 2018). Specifying an incorrect model for the response mechanism is easy, even for experts.

Most survey researchers do not claim to perfectly understand the response mechanism, but instead assume that they are able to estimate it well-enough in order to adjust for nonresponse. This is the final assumption that allows survey researchers to cling to the probability sampling paradigm.

In order for nonresponse to be missing at random, it must be possible to identify a set of features that, when controlled for, induces conditional independence between the response mechanism and outcome. In order for that to be true, we not only must correctly identify the auxiliary set, but then also observe each feature in both the survey and the population.

In practice, these two criteria are incredibly difficult to fulfill. For example, various studies have found strong evidence of differential partisan non-response, or large differences in response rates by political party affiliation (Kennedy et al., 2018; Gelman et al., 2016). However, there is no population data available on political partisanship available in many US states. Therefore, it is impossible to adjust for a known crucial member of the auxiliary set in many US states, like Michigan and Wisconsin. When this crucial population data is unavailable to researchers it is common to impute it, however that presents its own challenges.

While lack of education weighting is often cited as the main reason for the polling miss in the 2016 US presidential election, the post-mortem from American Association of Public Opinion Research (AAPOR) found that education only

explained part of the miss, and a large portion of the error was unexplained non-ignorable selection bias. In other words, missingness was MNAR.

Furthermore, response mechanisms are likely quite complex, however, in practice, researchers rely on a relatively small, consistent set of factors when adjusting for selection bias. For example, Pew Research used 12 variables in their weighting in its 2020 pre-election polling (Kennedy, 2020). The Cooperative Congressional Election Study conducted jointly by YouGov, Harvard, and a collection of survey researchers from across the United States, weighted its 2018 survey on 10 variables (Schaffner et al., 2019).

A more dire situation may arise if the response mechanism is directly dependent on an outcome of interest, and therefore one would need to include Y in $\mathbf{X}$ in order to satisfy $R \perp\!\!\!\perp Y|\mathbf{X}$, which is clearly impossible. No matter what data is or is not observed, selection bias is unrecoverable in this scenario. This could occur if partisanship directly contributes to how likely an individual is to answer a poll. For example, Donald Trump's consistent tendency to undermine institutions and claim that the 2020 election was "rigged" against him may have convinced his supporters to mistrust polls. This would indicate a direct dependence of response on election polls' main outcome of interest, a nonresponse mechanism which is impossible to recover from.

This specific response dynamic may not generalize to other survey contexts, yet highlights a case in which seemingly infinite resources, scrutiny, and expert insight are not always enough to prevent selection bias - some scenarios are truly unrecoverable.

Communication

Rather that introducing new selection bias when communicating results, the main threat that selection bias poses to the communication stage is in leading researchers to understate the true uncertainty associated with a set of results. This is not entirely the fault of researchers – we currently lack good analytical tools for quantifying uncertainty from sources other than random sampling. Shirani-Mehr et al. (2018) evaluates the first consequence by examining survey error in 4221 political polls in the US from 1998 to 2014, and finds that survey error was about twice as large as indicated by the margins of error reported with survey estimates.

When results are stated over-confidently, and then are wrong, consumers of polling data are left with an impression of "catastrophic bias," which occurred in both the 2016 and 2020 US presidential elections. In 2016, Nate Silver of FiveThirtyEight gave Hillary Clinton an 71% chance of winning the presidency, and the New York Times estimated that she had an 85% chance of winning the

election (Silver, 2016). FiveThirtyEight's forecast based on aggregated public polls predicted that Clinton would win 302 electoral college votes, including those of Wisconsin, Michigan, Pennsylvania, and Florida, all states that she ended up losing. While poll aggregation is powerful in that it leverages data from a wide variety of studies, it was unable to account for those studies' misunderstandings about the underlying response mechanism, including a systemic failure to recognize the importance of education weighting.

Election forecasters were more cautious in the 2020 US presidential election than in 2016. For example, FiveThirtyEight even added random noise to their state-level forecasts to increase the probability of rare-events (Gelman, 2020). However, the election results were much closer than anticipated, and even led to prominent Democratic pollsters admitting systemic error in their polling (Democracy Docket, 2021).

This phenomenon is closely related to the "replication crisis" that the broader field of applied statistics is currently grappling with. Breznau et al. (2022) demonstrates how many reputable researchers can arrive at vastly different conclusions when presented with the same research questions and the same data. Remarkably, Breznau et al. (2022) finds that "in spite of [a] highly granular decomposition of the analytical process we could only explain less than 4% of the variance in numerical outcomes and only 20% of the deviance in researchers' conclusions."

The replication crisis and the probability sampling crisis in some ways both stem from the failure of traditional uncertainty estimation tools, designed mainly to quantify sampling variability, to capture the additional uncertainty introduced by the study design itself. These latent sources of uncertainty can, as Breznau et al. (2022) demonstrates, actually contribute a great deal to the uncertainty in research conclusions.

1.3 Thesis Outline and Contributions

This thesis contains 4 papers, one of which has been published, followed by a conclusion. I also contributed to the following paper examining the impact of non-pharmaceutical interventions on the spread of COVID-19 in early 2020 in the United States. In order to present a cohesive set of work, it is not included in this thesis.

- Unwin, H.J.T., Mishra, S., **Bradley, V.C.** et al. State-level tracking of COVID-19 in the United States. Nat Commun 11, 6189 (2020). `https://doi.org/10.1038/s41467-020-19652-6`

Chapter 2 examines three surveys that all measure public opinion of COVID-19 vaccination from January-May 2021, but produce substantively different estimates of the rate of first-dose uptake among American adults. We compare survey estimates to a benchmark of first-dose uptake provided by the CDC and find that a surveys administered by Facebook and the Delphi Group ($n = 250,000$ per week) and the US Census Bureau ($n = 75,000$ per week) overestimate the benchmark by 17 and 14 percentage points in May 2021, respectively. Furthermore, due to the large sample size, the confidence intervals associated with these estimates are negligible. In contrast, a survey by Axios-Ipsos with about 1,000 respondents per week produces estimates that reliably track the CDC's benchmark and have well-calibrated confidence intervals. We use a framework for data quality (Meng, 2018) to decompose the error in these survey estimates and demonstrate that data quantity cannot efficiently compensate for low data quality (as measured by selection bias).

- Unrepresentative Big Surveys Significantly Overestimate US Vaccine Uptake **Bradley, V. C.***, Kuriwaki, S.*, Isakov, M., Sejdinovic, D., Meng, X. L., and Flaxman, S.*Nature*, 2021

We build on this work in Chapter 3, which introduces Active Learning Sampling Design (ALSD). ALSD uses ideas from Bayesian Optimization and Active Learning to propose a framework that integrates survey sampling and analysis. ALSD designs samples that can adapt to heterogeneous and dynamic nonresponse patterns, and that are tailored to analysis needs and modern analysis methods. We develop the mathematical framework for ALSD and demonstrate its advantages in simulation studies. Though it is not included in this paper, we have just completed a real-world test of ALSD with the World Food Programme in Zimbabwe and plan to present results in a subsequent paper.

- Active Learning Sampling Design (ALSD)
  Bradley, V. C., Semenova, E., Howes, A., Zhang, M., Rashid, T., Imai-Eaton, J., Sejdinovic, D., and Flaxman, S.

Chapter 4 interrogates one of the core assumptions of IPW adjustment, that population targets $p_{\mathbf{X}}$ of auxiliary variables used in weighting $\mathbf{X}$ are known for certain. In practice, $p_{\mathbf{X}}$ is often estimated from other surveys or modeled. For example, in US pre-election polling, pollsters aim to weight samples to reflect the characteristics of the electorate. However, voting is not compulsory so the exact composition of the electorate is unknown prior to election day, and instead, pollsters must estimate a

"likely voter universe" using data from prior similar elections, in-cycle indicators of enthusiasm, and contextual knowledge. The uncertainty in survey-based estimates of candidate vote share resulting from unknown population targets is, if not ignored entirely, rarely quantified. Here we introduce *leverage* as a tool for capturing some of that uncertainty. Leverage analytically captures the sensitivity of a weighted estimator of a population mean $\hat{Y}_w$ to error in population weighting targets. We derive leverage by decomposing the bias in $\hat{Y}_w$, introduce estimation methods, and assess their performance in simulation studies. We also demonstrate how leverage can be used to derive population frame uncertainty intervals.

- Leverage of weighting auxiliary variables
  Bradley, Valerie C., Schwenzfeier, M., and Sejdinovic, D.

Lastly, Chapter 5 examines selection bias in the UK Biobank (UKB) neurological imaging cohort, a nonprobability sample of approximately 30,000 participants. Largely written prior to Meng (2018), this paper begins to explore the idea that data quantity does not necessarily compensate for data quality. We begin by reviewing selection bias and nonresponse adjustment theory from the perspective of structural causal models (Pearl, 1995a). Then we introduce a new IPW adjustment technique that uses Bayesian Additive Regression Trees (BART) to aid with auxiliary variable selection, and test performance relative to raking, post-stratification, and other standard weighting techniques. We then demonstrate how these methods can be applied to the UKB neurological imaging cohort to adjust for known healthy volunteer bias and derive more accurate point estimates and estimates of associations, for example that of age and hippocampal volume.

- Addressing Selection Bias in the UK Biobank Neurological Imaging Cohort
  Bradley, V. C., Nichols, T. E.
  `https://www.medrxiv.org/content/10.1101/2022.01.13.22269266v2`

2

Unrepresentative Big Surveys Significantly Overestimate US Vaccine Uptake

Abstract

Accurate surveys are the primary tool for understanding public opinion towards and barriers preventing COVID-19 vaccine uptake. Conducting accurate surveys requires minimizing bias at every stage of the data collection process. These biases do not diminish with large sample sizes but rather are magnified by them, an instance of the Big Data Paradox (Meng, 2018). A clear demonstration comes from estimates of US vaccine uptake from the Delphi-Facebook (with about 250,000 responses per week) and Census Household Pulse (about 75,000 per week) surveys. Both significantly overestimate uptake compared to a benchmark from the Centers for Disease Control and Prevention (CDC)—by 17 and 14 percentage points respectively in May 2021. At the same time their large sample sizes lead (incorrectly) to negligible error bars. In contrast, a high-quality Axios-Ipsos panel (with about 1,000 responses) provides reliable estimates and error bars. We leverage a recently proposed framework for data quality (Meng, 2018) to decompose estimation error and to conduct a scenario analysis for implications on vaccine willingness and hesitancy. We show how a survey of 250,000 respondents can produce an estimate of a population mean that is no more accurate than an estimate from a simple random sample of size 10. Our main aim is to raise general awareness that compensating for low data quality by increasing data quantity is a mathematically provable losing proposition for assessing population averages, such as vaccination rates.

2.1 Which estimates should we trust?

Throughout the COVID-19 epidemic, public, timely and reliable datasets have played a crucial role in informing epidemic responses in government (Murthy et al., 2021) and civil society (Arrieta et al., 2021). The roll-out of vaccines across the US in 2021 has focused attention on critically important questions surrounding vaccine uptake, willingness, and hesitancy. Policymakers and the public urgently need fine-grained spatial, temporal, and sociodemographic information about COVID-19 vaccine related attitudes and behaviors (Murthy et al., 2021).

However, substantial discrepancies exist between two large sample-size surveys that measure vaccine-related behavior and attitudes in the US – Delphi-Facebook's COVID-19 symptom tracker (Barkay et al., 2020; Kreuter *et al.*, 2020) ($n \approx 250{,}000$ per week and with over 4.5 million responses from January to May 2021) and the Census Bureau's Household Pulse survey (Fields and Hunter-Childs *et al.*, 2020) ($n \approx 75{,}000$ per wave and with over 600,000 responses from January to May 2021) – and a more traditional probability-based online panel Axios-Ipsos' Coronavirus Tracker (Jackson et al., 2021) ($n \approx 1{,}000$ responses per wave, and over 10,000 responses from January to May 2021).

Under standard statistical assumptions, the large sample sizes of the first two surveys would yield negligible uncertainty intervals, making the divergences among these surveys' estimates all the more striking. For example, Delphi-Facebook state-level estimates for willingness to receive a vaccine from the end of March 2021 are 8.5 percentage points lower on average than those from the Census Household Pulse (Extended Data Fig. 2.3A), with differences as large as 16 percentage points. Such discrepancies can mislead, or at least confuse, policy-making.

While estimates of hesitancy and willingness must be derived from survey data, the US Centers for Disease Control and Prevention (CDC) compiles and reports vaccine uptake from administrative data sources. The availability of this data presents a unique benchmark through which to evaluate the reliability of survey estimates of vaccine uptake. Crucially, this is a valid comparison because none of the surveys use the CDC benchmark to adjust or assess estimates of vaccine uptake. The CDC has noted the discrepancies between their own reported vaccine uptake and that of the Census Household Pulse (Nguyen et al., 2021; Santibanez et al., 2021), and we find even larger discrepancies with the Delphi-Facebook data (Fig. 2.1a).

Even if estimates of absolute levels are wrong, one might hope that relative estimates, like changes in vaccine uptake, are correct. Unfortunately, errors have increased over time, from just a few percentage points in January 2021 to 4.2

percentage points (Axios-Ipsos), 14 percentage points (Census Household Pulse), and 17 percentage points (Delphi-Facebook) by mid-May (Fig. 2.1b). For context, for a state near the herd immunity threshold (70-80% based on recent estimates (Haas et al., 2021)), a discrepancy of 10 percentage points in vaccination rates could be the difference between containment and uncontrolled exponential growth in new SARS-CoV-2 infections.

Various research groups and the CDC provide spatially fine-grained estimates using these datasets (Institute for Health Metrics and Evaluation, 2021; Rader et al., 2022; US Centers for Disease Control and Prevention, 2021b), in keeping with the stated purpose of the large sample sizes. However, Extended Data Fig. 2.3G-H show that in March, Delphi-Facebook and Census Household Pulse over-estimated CDC state-level vaccine uptake by 16 and 9 percentage points, respectively. Relative estimates are again no better than absolute estimates: there is barely any agreement in a survey's estimated state-level rankings with the CDC (a Kendall rank correlation of 0.26 in Extended Data Fig. 2.3I, 0.21 in Extended Data Fig. 2.3J). For example, Massachusetts is ranked 45th and 41st (one of the lowest) in vaccine uptake by Delphi-Facebook and Census Household Pulse, respectively, but 5th (in the top ten) by the CDC.

2.2 The Big Data Paradox in estimating vaccine uptake

We focus on the Delphi-Facebook, and Census Household Pulse, surveys because their large sample sizes present the opportunity to examine the Big Data Paradox (Meng, 2018). We do not estimate vaccine uptake ourselves, but rather use estimates and standard errors provided in the data releases for each survey.

Delphi-Facebook and Census Household Pulse surveys persistently overestimate vaccine uptake relative to the CDC's benchmark (see Fig. 2.1a). Despite being the smallest survey by an order of magnitude, Axios-Ipsos' estimates track well the CDC rates, and their 95% confidence intervals contain the benchmark estimate from the CDC in 10 out of 11 surveys (an empirical coverage probability of 91%).

Bias in big surveys is particularly concerning because as sample size increases, bias (rather than variance) dominates estimator error. Conventional formulas for confidence intervals mislead by conveying dire overconfidence in biased estimates. Fig. 2.1a shows 95% confidence intervals for vaccine uptake based on reported sampling standard errors and weighting design effects (Kish, 1965). Axios-Ipsos has the widest confidence intervals, but also the smallest design effects (1.1-1.2)

suggesting that its accuracy is driven more by high-quality data collection rather than post-survey adjustment. Census Household Pulse has small, but visible, 95% confidence intervals that have been greatly inflated by large design effects (4.4-4.8) indicating large weighting adjustments; however, confidence intervals still fail to include the true rate of vaccine uptake. Most concerningly, confidence intervals for Delphi-Facebook are vanishingly small – driven by large sample size and moderate design effects (1.4-1.5) – indicating that samples are only moderately weighted, and that this adjustment is not nearly enough to correct for bias in data collection.

While it is well-understood that traditional confidence intervals only capture sampling errors in surveys (Groves et al., 2011) (and not the total errors), the traditional survey framework lacks analytic tools for quantifying nonsampling errors separately from sampling errors. In large surveys, sampling error becomes negligible, and error is dominated by unquantifiable sources.

When large samples are biased, they are therefore doubly misleading – they produce confidence intervals with incorrect centers and substantially underestimated widths. This is the Big Data Paradox (Meng, 2018): *the larger the data size, the surer we fool ourselves* when we fail to account for data quality.

2.3 A framework for analytically quantifying data quality

A recently proposed statistical framework (Meng, 2018) permits us to interrogate and quantify key sources of error in surveys, and hence address questions about conflicting survey estimates analytically. This framework has been applied to COVID case counts (Dempsey, 2020), and in other non-COVID settings (Isakov and Kuriwaki, 2020). Its full application requires ground-truth benchmarks, which are generally not available for most survey outcomes.

This data-quality framework consists of a data quality identity which decomposes the actual survey error in three terms:

$$\underbrace{\bar{Y}_n - \bar{Y}_N}_{\text{Actual Error}} = \underbrace{\hat{\rho}_{Y,R}}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{Data Quantity}} \times \underbrace{\sigma_Y}_{\text{Problem Difficulty}} . \tag{2.1}$$

This identity is explained and expressed in a more general form for weighted estimators in the Methods. Briefly, the actual error between the sample average $\bar{Y}_n$ (e.g., vaccination rate in a sample of size n) and the population average $\bar{Y}_N$ (e.g., vaccination rate in a population size of N) is determined by three factors, as listed in

the product form on the right-hand side of identity (2.1). The first factor, called *data defect correlation* (*ddc*) (Meng, 2018), is a measure of data quality by quantifying total bias (from any source), measured by the correlation between the recording indicator ($R = 1$ if an answer gets recorded and $R = 0$ otherwise) and the value recorded, Y. The second factor is a data quantity index which captures the impact of the sampling fraction $f = n/N$, emphasizing what matters is the relative sample size, not the absolute sample size n. The third factor reflects the problem difficulty by measuring the population heterogeneity (via standard deviation of Y), because the more heterogeneous is a population, the harder it is to estimate its average well.

Identity (2.1) allows us to calculate the bias-adjusted effective sample size n_{eff}, that is, the size of a simple random sample that we would expect to exhibit the same level of Mean Square Error as what was actually observed in a given study with a given *ddc*. Unlike the classical effective sample size (Kish, 1965), this quantity captures the impact of bias as well as that of variance increases from weighting and sampling. Details for this calculation are in Methods.

2.4 Decomposing survey error

While *ddc* is not directly observed, COVID-19 surveys present a rare case in which it can be deduced because all other terms in equation (2.1) are known: the sample size n of each survey wave, the estimate of vaccine uptake from each sample wave $\bar{Y}_{_N}$, and the population size N of US adults from US Census estimates (US Census Bureau, 2019). We use the CDC's report of the cumulative count of first doses administered to US adults as the benchmark (US Centers for Disease Control and Prevention, 2021a; Murthy et al., 2021), $\bar{Y}_{_N}$, and calculate $\sigma_{_Y} = \sqrt{\bar{Y}_{_N}\left(1 - \bar{Y}_{_N}\right)}$ because Y is binary (but identity (2.1) is not restricted to binary Y). We apply this framework to the aggregate error observed in Fig. 2.1a.

Our analysis relies on the accuracy of the underlying CDC benchmark, which may be subject to delays and slippage in how the CDC centralizes information from states (Tiu et al., 2022), and other systemic errors which administrative data can suffer from (Groen, 2012; Tu et al., 1993). As a sensitivity analysis to check the robustness of our findings to further misreporting, we present our results with sensitivity intervals under the assumption that CDC's reported numbers suffer from $\pm 5\%$ and $\pm 10\%$ error. These scenarios were chosen based on analysis of the magnitude by which the CDC's initial estimate for vaccine uptake by a particular day increases as the CDC receives delayed reports of vaccinations that occurred on that day (Supplementary Information 2.13.2). However, these scenarios may not

capture larger systemic issues affecting CDC vaccination reporting, issues requiring a systematic investigation of the quality of CDC's data themselves.

The error of each survey's estimate of vaccine uptake (Fig. 2.1b) increases over time for all studies, most markedly for Delphi-Facebook. Problem difficulty is a population quantity that changes over time and peaks when the true proportion is 50% (April 2021), then decreases again as the true proportion continues to rise above 50% (Fig. 2.1c). The data quantity index ($\sqrt{(N-n)/n}$) decreases with larger samples n and reflects that about 0.1%, 0.03%, and 0.0004% of the US adult population are sampled in each wave of Delphi-Facebook, Census Household Pulse and Axios-Ipsos, respectively (Fig. 2.1d).

The *ddc* increases over time for Census Household Pulse and, most significantly, for Delphi-Facebook (Fig. 2.1e). For Axios-Ipsos, it is much smaller and steady over time, consistent with what one would expect from a representative sample. This decomposition suggests that the increasing error in estimates of vaccine uptake in Delphi-Facebook and Census Household Pulse is primarily driven by increasing *ddc*, which captures the overall *impact* of the bias in coverage, selection, and response.

Small increases in *ddc* from what is expected in probability samples can result in drastic reductions of the bias-adjusted effective sample size, or the size of a simple random sample that would have the same expected mean squared error as what was actually observed in each survey. For estimating the US vaccination rate, Delphi-Facebook has a bias-adjusted effective sample size of less than 10 in April 2021, a 99.99% reduction from the average weekly sample size of 250,000 (Fig. 2.2). The Census Household Pulse also suffers from over 99% reductions in effective sample size by May 2021.

2.5 Comparing study designs

Understanding *why* bias occurs in some surveys but not others requires an understanding of the sampling strategy, modes, questionnaire, and weighting scheme of each survey. Table 2.1 compares the design of each survey (more details in the Methods section 2.9.7 and Extended Data Table 2.3).

Axios-Ipsos was designed to be representative of all US adults and Census Household Pulse was designed to rapidly measure how Americans' lives have been affected by the pandemic. Delphi-Facebook has stated that the intent of their survey is to make comparisons over space, time, and subgroups and that caution should be taken when using their data to make point estimates, like that of vaccine uptake.

However, we also note that Delphi-Facebook has reported point estimates of vaccine uptake in its own publications (The Delphi Group, 2021; Reinhart et al., 2021).

All three surveys are conducted online and target the US adult population, but vary in respondent recruitment methods (Kennedy et al., 2016). The Delphi-Facebook survey recruits respondents from active Facebook users (the Facebook Active User Base, or FAUB) using daily unequal-probability stratified random samples. The Census Bureau uses a systematic random sample to select households from the subset of the Census' Master Address File (MAF) for which they have obtained either phone or email contact information (approximately 81% of all households on the MAF).

In comparison, Axios-Ipsos relies on inverse response propensity sampling from Ipsos' (online) KnowledgePanel, which are participants Ipsos recruits from an address-based probabilistic sample from USPS's Delivery Sequence File (DSF). The DSF is similar to the Census' MAF. Unlike the Census Household Pulse, potential respondents are not limited to the subset for whom email and phone contact information is available. Furthermore, Ipsos provides internet access and tablets to recruited panelists who lack home internet access. In 2021, this "offline" group typically comprises 1% of the final survey.

All three surveys weight on age and gender, i.e. assign larger weights to respondents of underrepresented age-gender subgroups and smaller weights to those of overrepresented subgroups. Axios-Ipsos and Census Household Pulse also weight on education and race/ethnicity. And Axios-Ipsos additionally weights to the composition of political partisanship measured from the ABC News/Washington Post poll in 6 of the 11 waves we study. Education, a known correlate of propensity to respond to surveys (Kennedy et al., 2018) and social media use (Auxier and Anderson, 2021), are notably absent from Delphi-Facebook's weighting scheme, as is race/ethnicity. None of the surveys use the CDC benchmark to adjust or assess estimates of vaccine uptake.

2.6 Explanations for error

Table 2.2 illustrates some consequences of these design choices. Axios-Ipsos samples mimic the actual breakdown of education attainment among US adults even before weighting. After weighting, Axios-Ipsos and Census Household Pulse match the population benchmark, by design. Delphi-Facebook does not explicitly weight on education, and hence the education bias persists in their weighted estimates: those without a college degree are underrepresented by nearly 20 percentage points. The

story is similar for race/ethnicity. Delphi-Facebook's weighting scheme does not adjust for race/ethnicity, and hence their weighted sample still over-represents White adults by 8 percentage points, and under-represents Black and Asian proportions by around 50 percent of their size in the population.

This explains part of the error of Delphi-Facebook. The racial groups that Delphi-Facebook undersamples tend to be more willing and less vaccinated. In other words, re-weighting the Delphi-Facebook survey to upweight racial minorities may bring willingness estimates closer to Household Pulse and the vaccination rate closer to CDC. The three surveys also report that people without a 4-year college degree are less likely to have been vaccinated compared to those with a degree (Table 2.2 and Supplementary Information 2.15.1). If we assume that vaccination behaviors do not differ systematically between non-respondents and respondents *within* each demographic category, under-representation of less-vaccinated groups would contribute to the bias found here. However, this alone cannot explain the discrepancies in all the outcomes. Census Household Pulse weights on both race and education and still over-estimates vaccine uptake by over ten points in late May.

Delphi-Facebook and Census Household Pulse may be unrepresentative with respect to political partisanship, which has been found to be correlated with vaccine behavior (Gadarian et al., 2021) and with survey response (Mercer et al., 2018), and thus may contribute to observed bias. However, neither Delphi-Facebook nor Census Household Pulse collects partisanship of respondents; Census agencies are prohibited from asking about political preference. Moreover, no unequivocal population benchmark for partisanship exists.

Rurality may also contribute to the errors as it correlates with vaccine status (Murthy et al., 2021) and home internet access (Ryan, 2017). Neither the Census Household Pulse nor Delphi-Facebook weights on sub-state geography, which may mean that adults in more rural areas are less likely to be vaccinated and also underrepresented in the surveys, leading to overestimation of vaccine uptake.

Axios-Ipsos weights to metropolitan status and also recruits a fraction of its panelists from an "offline" population of individuals without Internet access. We find that *dropping* these offline respondents ($n = 21$, or 1 percent of the sample) in their March 22 wave *increases* Axios-Ipsos' overall estimate of the vaccination rate by 0.5 percentage points, thereby increasing the total error. The offline population is simply too small to explain the entirety of the difference in accuracy between Axios-Ipsos and either the Census Household Pulse (6 percentage points) or Delphi-Facebook (14 percentage points), in this time period.

Absent the full set of weighting variables and population targets, careful recruitment of panelists is at least as important as weighting. Weighting on observed covariates alone cannot explain or correct the discrepancies we observe. For example, reweighting Axios-Ipsos survey data using only Delphi-Facebook's weighting variables (age group and gender), increased the error in their vaccination estimates by 1 percentage point, but this estimate with Axios-Ipsos data is still more accurate than that with Delphi-Facebook data. The Axios-Ipsos estimate with Delphi-Facebook weighting overestimated vaccination by 2 percentage points, whereas Delphi-Facebook overestimated it by 11 percentage points.

The implication is that there is no silver bullet: every small part of panel recruitment, sampling, and weighting matters for controlling the *ddc*. In multi-stage sampling, which includes for instance the selection of participants followed by non-response, even a *single* step involving a non-representative sample can substantially bias the final result (see Supplementary Information 2.9.5). A *total quality control* approach — inspired by the Total Survey Error framework (Biemer and Lyberg, 2003) — is a better strategy than trying to prioritize some components over others in order to improve data quality.

2.7 Addressing common misperceptions

The three surveys discussed in this article demonstrate a seemingly paradoxical phenomenon – the two "big surveys" are far more confident, yet also far more biased, than the smaller, more traditional Axios-Ipsos poll. Our findings are paradoxical only when we fall into the trap of the long-held, but incorrect, intuition that estimation errors necessarily decrease as we collect more data (Mayer-Schönberger and Cukier, 2013).

A limitation of our vaccine uptake analysis is that it focuses on a specific outcome for which a benchmark is available. The *ddc* is defined with respect to a specific outcome, vaccination uptake, yet bias in this outcome does not mean that estimates for all outcomes (e.g., vaccine hesitancy) suffer from the same total error (Groves and Peytcheva, 2008). However, we caution this optimism when there is reason to believe that bias is caused by unrepresentativeness of the sample. We observe under-representation of low-education and non-white respondents in the Census Household Pulse raw data and Delphi-Facebook's raw and weighted data, and neither survey is weighted on behavioral characteristics, all of which have been linked to bias in a range of survey outcomes (Kennedy et al., 2018; Mercer et al., 2018). In addition, we defined the outcomes so that hesitancy, willingness, and

vaccination sum to 100%, so if vaccination is mismeasured one or both of hesitancy and willingness must be incorrect as well (Extended Data Fig. 2.6). More broadly, it is likely that the *ddc* for variables correlated with vaccination uptake are of similar magnitudes. Given that vaccine hesitancy and COVID-19-related attitudes are the main focus of these surveys, the outcome studied here may implicate the accuracy of many other variables. Finally, we find that, for vaccine uptake, bias is not limited to population point estimates, but also affects estimates of changes over time (contrary to published guidance (Kreuter *et al.*, 2020)) – both Delphi-Facebook and Census Household Pulse significantly overestimate the slope of vaccine uptake over time relative to that of the CDC benchmark (Fig. 2.1b).

Some may argue that bias is a necessary trade-off for having data that is sufficiently large for conducting highly granular analysis, such as county-level estimation of vaccine hesitancy (US Centers for Disease Control and Prevention, 2021b). While high-resolution inference is important, we caution that this is a double-edged argument. A highly biased estimate with a misleadingly small confidence interval can do more damage than having no estimate at all.

The accuracy of our analysis also relies on the accuracy of the CDC's estimates of COVID vaccine uptake. However, if the selection bias in the CDC's benchmark is significant enough to alter our results, then that itself would be yet another example of the Big Data Paradox.

2.8 Discussion

This is not the first time that the Big Data Paradox has reared its head: Google Trends predicted more than twice the number of influenza-like illnesses than the CDC in February 2013 (Lazer et al., 2014). Though the studies we consider were more carefully designed than Google Trends, they are still susceptible to similar biases. Delphi-Facebook is "the largest public health survey ever conducted in the United States" (Salomon et al., 2021). The Census Household Pulse is conducted in collaboration between the US Census Bureau and eleven statistical government partners, all with enormous resources and survey expertise. Both studies take steps to mitigate selection bias, yet overestimate vaccine uptake by double digits. As we demonstrated, the impact of bias is magnified as relative sample size increases.

In contrast, Axios-Ipsos records only about 1,000 responses per wave, but makes additional efforts to prevent selection bias. Small surveys could be just as wrong as large surveys in expectation – of the three other small to medium online surveys additionally analyzed, two also miss the CDC vaccination benchmark

(Extended Data Fig. 2.7). The overall lesson is that investing in data quality (particularly during collection, but also in analysis) minimizes error more efficiently than does increasing data quantity. Of course, a sample size of 1,000 may be too small (i.e. leading to unhelpfully large confidence intervals) for the kind of 50-state estimates given by big surveys. However, small area methods that borrow information across subgroups (Park et al., 2004) can perform better with better quality, albeit small, data, and it is an open question whether that approach would outperform the large, biased surveys.

There are approaches to correct for these bias in both probability and non-probability samples alike. For COVID-19 surveys in particular, since June 2021, the AP-NORC multi-mode panel has weighted their COVID-19 related surveys to the CDC benchmark, so that the weighted *ddc* for vaccine uptake is zero by design (Associated Press-NORC Center for Public Affairs Research, 2021). More generally, there is an extensive literature on approaches for making inferences from data collected from nonprobability samples (Wang et al., 2015; Elliott and Valliant, 2017; Little et al., 2020). Other promising approaches include integrating surveys of varying quality (Wiśniowski et al., 2020; Yang et al., 2020), and leveraging the estimated *ddc* in one outcome to correct bias in others under several scenarios (Supplementary Information 2.16).

While more needs to be done to fully examine the nuances of large surveys, organically collected administrative datasets, and social media data, we hope this first comparative study of *ddc* highlights the alarming implications of the *Big Data Paradox* – how large sample sizes magnify the impact of even small defects in data collection, leading to overconfidence in incorrect inferences.

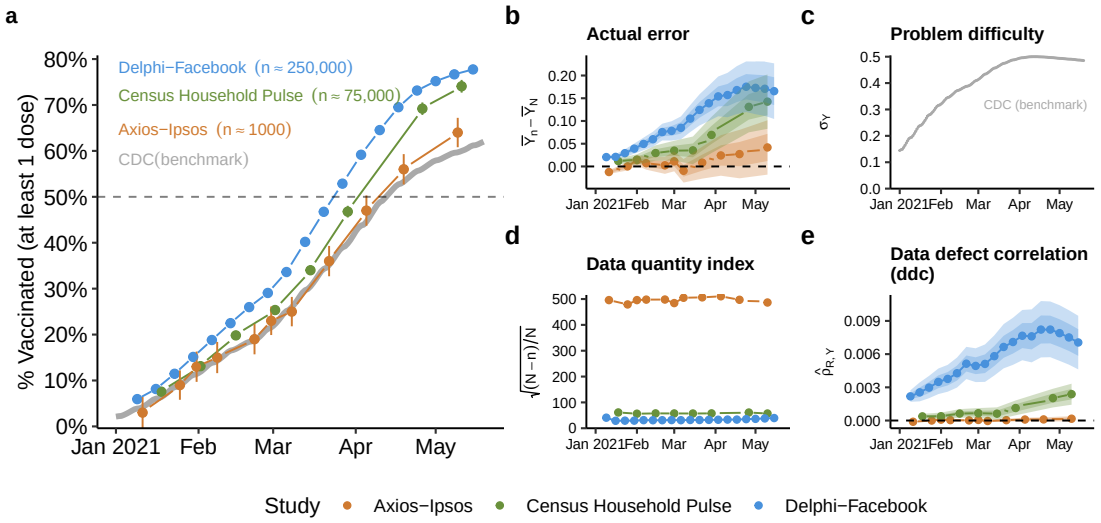


Figure 2.1: Conflicting estimates of vaccine uptake. **a.** Estimates of vaccine uptake for US adults in 2021 compared to CDC benchmark data, plotted by end date of each survey wave. Points indicate each study's estimate of vaccine uptake, and intervals are 95% CIs using reported standard errors and design effects. Delphi-Facebook has $n = 4,525,633$ across 19 waves, Census Household Pulse has $n = 606,615$ across 8 waves, and Axios-Ipsos has $n = 11,421$ across 11 waves. Delphi-Facebook's CIs are too small to be visible. **b.** Total error $\bar{Y}_n - \bar{Y}_N$, **c.** problem difficulty σ_Y, **d.** index of data quantity $\sqrt{(N-n)/n}$, **e.** data defect correlation. Shaded bands represent scenarios of +/-5% (darker) and +/-10% (lighter) error in CDC benchmark relative to reported values (points). **b** - **d** comprise the decomposition in Equation 2.1.

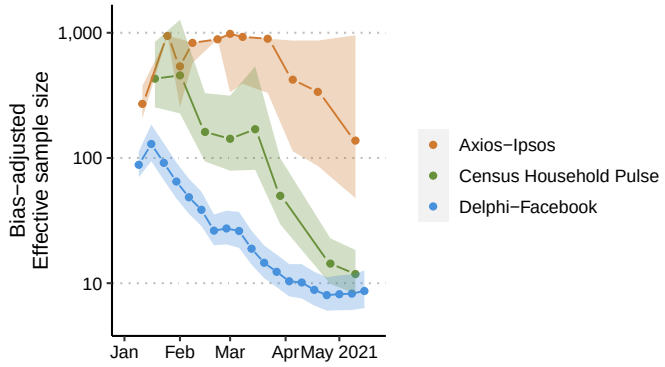


Figure 2.2: Bias-adjusted effective sample size. The bias-adjusted effective sample size of an estimate (different from the classic Kish effective sample size) is the sample size of a simple random sample which would have the same Mean Square Error of the estimate. Effective sample sizes shown on the $\log_{10}$ scale. The original sample size was $n = 4{,}525{,}633$ across 19 waves for Delphi-Facebook, $n = 606{,}615$ across 8 waves for Census Household Pulse, $n = 11{,}421$ across 11 waves for Axios-Ipsos. Shaded bands represent scenarios of +/-5% error in the CDC benchmark relative to point estimates based on actual reported values.

	Axios-Ipsos	Census Household Pulse	Delphi-Facebook
Recruitment mode	Address-based mail sample to Ipsos KnowledgePanel	SMS and email	Facebook Newsfeed
Interview mode	Online	Online	Online
Average size	1,000/wave	75,000/wave	250,000/week
Sampling frame	Ipsos KnowledgePanel; internet/tablets provided to ∼5% of panelists who lack home internet	Census Bureau's Master Address File (individuals for whom email / phone contact information is available)	Facebook active users
Weighting variables	Gender by age, race, education, Census region, metropolitan status, household income, partisanship.	Education by age by sex by state, race/ethnicity by age by sex by state, household size	Stage 1: age, gender "other attributes which we have found in the past to correlate with survey outcomes" to FAUB; Stage 2: state by age by gender

Table 2.1: Comparison of survey designs. The Table compares key design choices across Axios-Ipsos, Census Household Pulse, and Delphi-Facebook studies. All surveys target the US adult population. See Extended Data Table 2.3 for additional comparisons.

	Composition of US Adults							Survey Estimates		
	Axios-Ipsos		Household Pulse		Delphi-Facebook		ACS	Household Pulse		
	Raw	Weighted	Raw	Weighted	Raw	Weighted	Benchmark	Vax	Will	Hes
Education										
High School	35%	39%	14%	39%	19%	21%	39%	39%	40%	21%
Some College	29	30	32	30	36	36	30	44	38	18
4-Year College	19	17	29	17	25	25	19	54	36	10
Post-Graduate	17	14	26	13	20	18	11	67	26	7
Race/Ethnicity										
White	71%	63%	75%	62%	74%	68%	60%	50%	33%	17%
Black	10	12	7	11	6	6	12	42	39	19
Hispanic	11	16	10	17	11	16	16	38	48	14
Asian			5	5	2	3	6	51	43	5

Table 2.2: Composition of survey respondents by educational attainment and race/ethnicity. Axios-Ipsos: wave ending March 22, 2021, $n = 995$. Census Household Pulse: wave ending March 29, 2021, $n = 76{,}068$. Delphi-Facebook: wave ending March 27, 2021, $n = 181{,}949$. Benchmark uses the 2019 US Census American Community Survey (ACS), composed of roughly 3 million responses. Right-most column shows estimates of vaccine uptake (Vax), willingness (Will) and hesitancy (Hes) from the Census Household Pulse of the same wave.

2.9 Appendix: Methods

2.9.1 Calculation and interpretation of *ddc*

The mathematical expression for Equation (2.1) is given here for completeness:

$$\bar{Y}_n - \bar{Y}_N = \hat{\rho}_{Y,R} \times \sqrt{\frac{N-n}{n}} \times \sigma_Y \tag{2.2}$$

The first factor $\hat{\rho}_{Y,R}$ is called the *data defect correlation*, *ddc*, (Meng, 2018). It is a measure of data quality represented by the correlation between the recording indicator R ($R = 1$ if an answer is recorded and $R = 0$ otherwise) and its value, Y. Given a benchmark, the *ddc* $\hat{\rho}_{Y,R}$ can be calculated by substituting known quantities into Equation (2.2). In the case of a single survey wave of a COVID-19 survey, n is the sample size of the survey wave, N is the population size of US adults from US Census estimates (US Census Bureau, 2019), $\bar{Y}_n$ is the survey estimate of vaccine uptake, and $\bar{Y}_N$ is the estimate of vaccine uptake for the corresponding period taken from the CDC's report of the cumulative count of first doses administered to US adults (US Centers for Disease Control and Prevention, 2021a; Murthy et al., 2021). We calculate $\sigma_Y = \sqrt{\bar{Y}_N \left(1 - \bar{Y}_N\right)}$ because Y is binary (but Equation (2.2) is not restricted to binary Y).

We calculate $\hat{\rho}_{Y,R}$ by using *total* error $\bar{Y}_n - \bar{Y}_N$, which captures not only selection bias but also any measurement bias (e.g., from question wording). However, with this calculation method, $\hat{\rho}_{Y,R}$ lacks the direct interpretation as a correlation between Y and R, and instead becomes a more general index of data quality directly related to classical design effects (see Methods section "Bias-adjusted effective sample size").

It is important to point out that the increase in *ddc* does not necessarily imply that the response mechanisms for Delphi-Facebook and Census Household Pulse have changed over time. The correlation between a changing *outcome* and a steady response mechanism could change over time, hence changing the value of *ddc*. For example, as more individuals become vaccinated, and vaccination status is driven by individual behavior rather than eligibility, the correlation between vaccination status and propensity to respond could increase even if propensity to respond for a given individual is constant. This would lead to large values of *ddc* over time, reflecting the *increased impact* of the same response mechanism.

2.9.2 Error decomposition with survey weights

The data quality framework given by Equations (2.1) and (2.2) is a special case of a more general framework for assessing the actual error of a weighted estimator $\bar{Y}_w = \sum_i w_i R_i Y_i / \sum_i w_i R_i$, where w_i is the survey weight assigned to individual i. It is shown in Meng (2018) that

$$\bar{Y}_{\rm w} - \bar{Y}_{N} = \hat{\rho}_{Y,R_{\rm w}} \times \sqrt{\frac{N - n_{\rm w}}{n_{\rm w}}} \times \sigma_{Y}, \tag{2.3}$$

where $\hat{\rho}_{Y,R_{\rm w}} = {\rm Corr}(Y, R_{\rm w})$ is the finite population correlation between Y_i and $R_{{\rm w},i} = w_i R_i$ (over $i = 1, \ldots, N$). The "hat" on ρ reminds us that this correlation depends on the specific realization of $\{R_i, i = 1, \ldots, N\}$. The term $n_{\rm w}$ is the classical "effective sample size" due to weighting (Kish, 1965), i.e., $n_{\rm w} = n/(1 + CV_{\rm w}^2)$, where $CV_{\rm w}$ is the coefficient of variation of the weights for all individuals in the observed sample, that is, the standard deviation of weights normalized by their mean. It is common for surveys to rescale their weights to have mean 1, in which case CV_w^2 is simply the sample variance of W.

When all weights are the same, Equation (2.3) reduces to Equation (2.2). In other words, the *ddc* term $\hat{\rho}_{Y,R_{\rm w}}$ now also takes into account the impact of the weights as a means to combat the selection bias represented by the recording indicator R. Intuitively, if $\hat{\rho}_{Y,R} = {\rm Corr}(Y, R)$ is high (in magnitude), then some Y_i's have a higher chance of entering our data set than others, thus leading to a sample average that is a biased estimator for the population average. Incorporating appropriate weights can reduce $\hat{\rho}_{Y,R}$ to $\hat{\rho}_{Y,R_{\rm w}}$, with the aim to reduce the impact of the selection bias. However, this reduction alone may not be sufficient to improve the accuracy of $\bar{Y}_w$ because the use of weight necessarily reduces the sampling fraction $f = n/N$ to $f_w = n_w/N$ as well since $n_w < n$. Equation (2.3) precisely describes this trade off, providing a formula to assess when the reduction of *ddc* is significant to outweigh the reduction of the effective sample size.

Measuring the correlation between Y and R is not a new idea in survey statistics (though note that *ddc* is the population correlation between Y and R, not the sample correlation), nor is the observation that as sample size increases, error is dominated by bias instead of variance (Bethlehem, 2002; Meng, 2014). The new insight is that *ddc* is a general metric to index the *lack of* representativeness of the data we observe, regardless of whether or not the sample is obtained via a probabilistic scheme, or weighted to mimic a probabilistic sample. As discussed in the section on addressing common misconception, any single *ddc* deviating from what is expected under representative sampling (e.g., probabilistic sampling) is

sufficient to establish the sample is not representative (but the converse is not true). Furthermore, the *ddc* framework refutes the common belief that increasing sample size necessarily improves statistical estimation (Meng and Xie, 2014; Meng, 2018).

2.9.3 Bias-adjusted effective sample size

By matching the mean-squared error of $\bar{Y}_w$ with the variance of the sample average from simple random sampling, Meng (2018) derives the following formula for calculating a *bias-adjusted effective sample size*, or n_{eff}:

$$n_{\text{eff}} = \frac{n_{\text{w}}}{N - n_{\text{w}}} \times \frac{1}{E[\hat{\rho}^2_{Y,R_{\text{w}}}]}$$

Given an estimator $\bar{Y}_{\text{w}}$ with expected total Mean Squared Error (MSE) T due to data defect, sampling variability, and weighting, this quantity n_{eff} represents the size of a simple random sample such that its mean $\bar{Y}_{N}$, as an estimator for the same population mean $\bar{Y}_{N}$, would have the identical MSE T. The term $E[\hat{\rho}^2_{Y,R_{\text{w}}}]$ represents the amount of selection bias (squared) expected on average from a particular recording mechanism R and a chosen weighting scheme.

For each survey wave, we use $\hat{\rho}^2_{Y,R_{\text{w}}}$ to approximate $E[\hat{\rho}^2_{Y,R_{\text{w}}}]$. This estimation is unbiased by design, since we use an estimator to estimate its expectation. Therefore, the only source of error is the sampling variation, which is typically negligible for large surveys, such as for Delphi-Facebook and the Census Household Pulse surveys. This estimation error may have more impact for smaller traditional surveys, such as Axios-Ipsos' survey, an issue we will investigate in subsequent work.

We compute $\hat{\rho}_{Y,R_{\text{w}}}$ by using the benchmark $\bar{Y}_{N}$, namely, via solving Equation (2.3) for $\hat{\rho}_{Y,R_{\text{w}}}$,

$$\hat{\rho}_{Y,R_{\text{w}}} = \frac{Z_w}{\sqrt{N}}, \quad \text{where} \quad Z_w = \frac{\bar{Y}_{\text{w}} - \bar{Y}_{N}}{\sqrt{\frac{1-f_{\text{w}}}{n_{\text{w}}}}\sigma_Y}. \tag{2.4}$$

We introduce this notation Z_w because it is the quantity that determines the well-known survey efficiency measure, the so-called *design effect*, which is the variance of Z_w for a probabilistic sampling design (Kish, 1965) (when we assume the weights are fixed). For the more general setting where $\bar{Y}_w$ may be biased, we replace the variance by MSE, and hence the bias-adjusted design effect $D_e = E[Z_w^2]$, which is the MSE relative to the benchmark measured in the unit of the variance of an average from a simple random sample of size n_w. Hence $D_I \equiv E[\hat{\rho}^2_{Y,R_{\text{w}}}]$, which was termed as the *data defect index* (Meng, 2018), is simply the bias-adjusted design effect *per unit*, because $D_I = D_e/N$.

Furthermore, because Z_w is the standardized actual error, it captures any kind of error inherited in $\bar{Y}_w$. This observation is important because when Y is subject to measurement errors, $Z_w/\sqrt{N}$ no longer has the simple interpretation as a correlation. But because we estimate D_I by Z_w^2/N directly, our effective sample size calculation is still valid even when Equation (2.3) does not hold.

2.9.4 Asymptotic behavior of *ddc*

As shown in Meng (2018), for any probabilistic sample without selection biases, the *ddc* is on the order of $1/\sqrt{N}$. Hence the magnitude of $\hat{\rho}_{Y,R}$ (or $\hat{\rho}_{Y,R_w}$) is small enough to cancel out the impact of $\sqrt{N-n}$ (or $\sqrt{N-n_w}$) in the data scarcity term on the actual error, as seen in Equation (2.2) (or Equation (2.3)). However, when a sample is unrepresentative, e.g. when those with $Y=1$ are more likely to enter the dataset than those with $Y=0$, then $\hat{\rho}_{Y,R}$ can far exceed $1/\sqrt{N}$ in magnitude. In this case, error will increase with $\sqrt{N}$ for a fixed *ddc* and growing population size N (Equation (2.2)). This result may be counter-intuitive in the traditional survey statistics framework, which often considers how error changes as sample size n grows. The *ddc* framework considers a more general setup, taking into account individual response behavior, including its impact on sample size itself.

As an example of how response behavior can shape both total error and the number of respondents n, suppose individual response behavior is captured by a logistic regression model

$$\text{logit}[\Pr(R=1|Y)] = \alpha + \beta Y. \tag{2.5}$$

This is a model for a response propensity score. Its value is determined by α, which drives the overall sampling fraction $f = n/N$, and by β, which controls how strongly Y influences whether a participant will respond or not.

In this logit response model, when $\beta \neq 0$, $\hat{\rho}_{Y,R}$ is determined by individual behavior, not by population size N. In Supplementary Information 2.14.1, we prove that *ddc* cannot vanish as N grows, nor can the observed sample size n ever approach 0 or N for a given set of (finite and plausible) values of $\{\alpha, \beta\}$, because there will always be a non-trivial percentage of non-respondents. For example, an f of 0.01 can be obtained under this model for either $\alpha = -0.46, \beta = 0$ (no influence of individual behavior on response propensity), or for $\alpha = -3.9, \beta = -4.84$. However, despite the same f, the implied *ddc* and consequently the MSE will differ. For example, the MSE for the former (no correlation with Y) is 0.0004, while the MSE for the latter (a -4.84 coefficient on Y) is 0.242, over 600 times larger.

See Supplementary Information 2.14.2 for the connection between *ddc* and a well-studied non-response model from econometrics, the Heckman selection model (Heckman, 1979).

2.9.5 Population size in multi-stage sampling

We have shown that the asymptotic behavior of error depends on whether the data collection process is driven by individual response behavior or by survey design. The reality is often a mix of both. Consequently, the relevant "population size" N depends on when and where the representativeness of the sample is destroyed, i.e., when the individual response behaviors come into play. Real-world surveys that are as complex as the three surveys we analyze here have multiple stages of sample selection.

Extended Data Table 2.5 takes as an example the sampling stages of the Census Household Pulse, which has the most extensive set of documentation among the three surveys we analyze. As we have summarized (Table 2.1 and Extended Data Table 2.3), the Census Household Pulse (1) first defines the sampling frame as the reachable subset of the Master Address File, (2) takes a random sample of that population to prompt (send a survey questionnaire), and (3) waits for individuals to respond to that survey. Each of these stages reduces the desired data size, and the corresponding *population size* is the intended sample size from the prior stage (in notation, $N_s = n_{s-1}$, for $s = 2, 3$). For example, for stage 3, the population size N_3 is the size of the intended sample size n_2 from the second stage, i.e., the sampling stage, because only the sampled individuals have a chance to respond.

Although all stages contribute to the grand *ddc*, the stage that dominates is the *first stage at which the representativeness of our sample is destroyed*— whose size will be labeled as the *dominating population size (dps)*—when the relevant population size decreases dramatically at each step. However, we must bear in mind that *dps* refers to the worst case scenario, when biases accumulate, instead of (accidentally) canceling each other out.

For example, if the 20 percent of the MAF excluded from the Census Household Pulse sampling frame (because they had no cell phone or email contact information) is not representative of the US Adult population, then the *dps* is N_1, or 255 million adults contained in 144 million households. Then the increase in bias for given *ddc* is driven by the rate of $\sqrt{N_1}$ where $N_1 = 2.55 \times 10^8$ and is large indeed (with $\sqrt{2.5 \times 10^8} \approx 15,000$). In contrast, if the sampling frame is representative of the target population and the outreach list is representative of the frame (and hence representative of the US adult population) but there is non-response bias, then *dps* is $N_3 = 10^6$ and the impact *ddc* is amplified by the square root of that number ($\sqrt{10^6} = 1,000$). In contrast, Axios-Ipsos reports a response rate of about 50%, and obtains a sample of $n = 1000$, so the *dps* could be as small as $N_3 = 2000$ (with $\sqrt{2000} \approx 45$).

This decomposition is why our comparison of the surveys is consistent with the *Law of Large Populations* (estimation error increases with $\sqrt{N}$), *even though all three surveys ultimately target the same US Adult Population.* Given our existing knowledge about online-offline populations (Ryan, 2017) and our analysis of Axios-Ipsos' small "offline" population, Census Household Pulse may suffer from unrepresentativeness at Stage 1 of Extended Data Table 2.5 where $N = 255$ million, and Delphi-Facebook may suffer from unrepresentativeness at the initial stage of starting from the Facebook User Base. In contrast, the main source of unrepresentativeness for Axios-Ipsos maybe at a later stage where the relevant population size is orders of magnitude smaller.

2.9.6 CDC estimates of vaccination rates

The CDC benchmark data used in our analysis was downloaded from the CDC's COVID data tracker (US Centers for Disease Control and Prevention, 2021a). We employ the cumulative count of people who have received at least one dose of COVID-19 vaccine reported in the "Vaccination Trends" tab. This data set contains vaccine uptake counts for all US residents (not only adults). However, the surveys of interest estimate vaccine uptake among adults. The CDC receives age-group-specific data on vaccine uptake from all states except for Texas on a daily basis, which is also reported cumulatively over time.

Therefore, we must impute the number of adults who have received at least one dose on each day. We assume Texas is exchangeable with the rest of the states in terms of the age distribution for vaccine uptake. Under this assumption, for each day, we use the age group vaccine uptake data from all states except for Texas to calculate the proportion of cumulative vaccine recipients who are 18 or older, then we multiply that number by the total number of people who have had at least one dose to estimate the number of US *adults* who have received at least one dose.

The CDC performs a similar imputation for the 18+ numbers reported in their COVID data tracker. However the CDC's imputed 18+ number is available only as a snapshot and not a historical time series, hence the need for our imputation. See Supplementary Information for details of the imputation implementation.

2.9.7 Additional survey methodology

The Census Household Pulse and Delphi-Facebook surveys are the first of their kind for each organization, while Ipsos has maintained their online panel for 12 years.

Question wording All three surveys ask whether respondents have received a COVID-19 vaccine. See Extended Data Table 2.3. Delphi-Facebook and Census Household Pulse ask similar questions ("Have you had / received a COVID-19 vaccination / vaccine?"). Axios-Ipsos asks "Do you personally know anyone who has already received the COVID-19 vaccine?," and respondents are given response options including "Yes, I have received the vaccine." The Axios-Ipsos question wording might pressure respondents to conform to their communities' modal behavior and thus misreport their true vaccination status, or may induce acquiescence bias from the multiple "yes" options presented. This pressure may exist both in high- and low-vaccination communities, so its net impact on Axios-Ipsos' results is unclear. Nonetheless, Axios-Ipsos' question wording does differ from that of the other two surveys, and may contribute the observed differences in estimates of vaccine uptake across surveys.

Population of Interest All three surveys target US adult population, but with different sampling and weighting schemes. Household Pulse sets the denominator of their percentages as the household civilian, non-institutionalized population in the United States of 18 years of age or older, excluding Puerto Rico or the island areas. Axios-Ipsos designs samples to be representative of the US general adult population 18 or older. For Facebook, the US target population reported in weekly contingency tables is the US adult population, excluding Puerto Rico and other US territories. For the CDC Benchmark, we define the denominator as the US 18+ population, excluding Puerto Rico and other US territories. To estimate the size of the total US population, we use the US Census Bureau Annual Estimates of the Resident Population for the United States and Puerto Rico, 2019 (US Census Bureau, 2019). This is also what the CDC uses as the denominator in calculating rates and percentages of the US population (US Centers for Disease Control and Prevention, 2021d).

Axios-Ipsos and Delphi-Facebook generate target distributions of the US adult population using the Current Population Survey (CPS), March Supplement, from 2019 and 2018, respectively. Census Household Pulse uses a combination of 2018 1-year American Community Survey (ACS) estimates and the Census Bureau's Population Estimates Program (PEP) from July 2020. Both the CPS and ACS are well-established large surveys by the Census and the choice between them is largely inconsequential.

Axios-Ipsos Data The Axios-Ipsos Coronavirus tracker is an ongoing, bi-weekly tracker intended to measure attitudes towards COVID-19 of adults in the US. The tracker has been running since March 13, 2020 and has released results from 45 waves as of May 28, 2021. Each wave generally runs over a period of 4 days. The Axios-Ipsos data used in this analysis was scraped from the topline PDF reports released on the Ipsos website (Jackson et al., 2021). The PDF reports also contain Ipsos' design effects, which we have confirmed are calculated as 1 plus the variance of the (scaled) weights.

Census Household Pulse Data The Census Household Pulse is an experimental product of the US Census Bureau in collaboration with eleven other federal statistical agencies. We use the point estimates presented in Data Tables, as well as the standard errors calculated by the Census Bureau using replicate weights. The design effects are not reported, however we can calculate it as $1 + CV_{\text{w}}^2$, where CV_{w} is the coefficient of variation of the individual-level weights included in the microdata (Kish, 1965).

Delphi-Facebook COVID symptom survey The Delphi-Facebook COVID symptom survey is an ongoing survey collaboration between Facebook, the Delphi Group at Carnegie Mellon University (CMU), and the University of Maryland (Barkay et al., 2020). The survey is intended to track COVID-like symptoms over time in the US and in over 200 countries. We use only the US data in this analysis. The study recruits respondents using a daily stratified random samples recruiting a cross-section of Facebook Active Users. New respondents are obtained each day, and aggregates are reported publicly on weekly and monthly frequencies. The Delphi-Facebook data used here was downloaded directly from CMU's repository for weekly contingency tables with point estimates and standard errors.

Data availability Raw data is deposited in the Harvard Dataverse `https://doi.org/10.7910/DVN/GKBUUK`. Data was collected from publicly available repositories of survey data by downloading it directly or using APIs.

Code availability Code to replicate the findings is available in the repository `https://github.com/vcbradley/ddc-vaccine-US`.

2.10 Appendix: Ethical compliance

According to HRA decision tools (http://www.hra-decisiontools.org.uk/research/), our study is considered Research, and according to the NHS REC review tool (http://www.hra-decisiontools.org.uk/ethics/), we do not need NHS Research Ethics Committee (REC) review, as we only used (1) publicly available, (2) anonymized, and (3) aggregated data outside of clinical settings.

2.11 Appendix: Acknowledgments

We thank Frauke Kreuter, Alex Reinhart, and the Delphi Group at Carnegie Mellon University, Facebook's Demography and Survey Science group; Frances Barlas, Chris Jackson, Mallory Newall, and the Public Affairs team at Ipsos; and Jason Fields and Jennifer Hunter Childs at the US Census Bureau for productive conversations about their surveys. We further thank the Delphi Group at CMU for their help in computing weekly design effects for the Delphi-Facebook COVID symptom survey, and the Ipsos team for providing flags for their "offline" respondents. We thank the US Centers for Disease Control and Prevention for responding to our questions, as well as Susan Paddock and other participants at the JPSM 2021 lecture (delivered by Meng) for their comments. We thank the anonymous reviewers for their constructive and helpful comments, which substantially improved our work. Finally, we thank Ariel Edwards-Levy for a tweet which originally inspired our interest in this topic.

Appendix: Funding V.B. is funded by the University of Oxford's Clarendon Fund and the EPSRC and MRC through the OxWaSP CDT programme (EP/L016710/1). X-L. M acknowledges partial financial support by NSF. S.F. acknowledges the support of the EPSRC (EP/V002910/1).

Appendix: Author contributions V.B. and S.F. conceived and formulated the research questions. All authors contributed to methodology, writing, visualization, editing, and data analysis.

Appendix: Competing Interests Authors have no competing interests.

2.12 Appendix: Extended Data

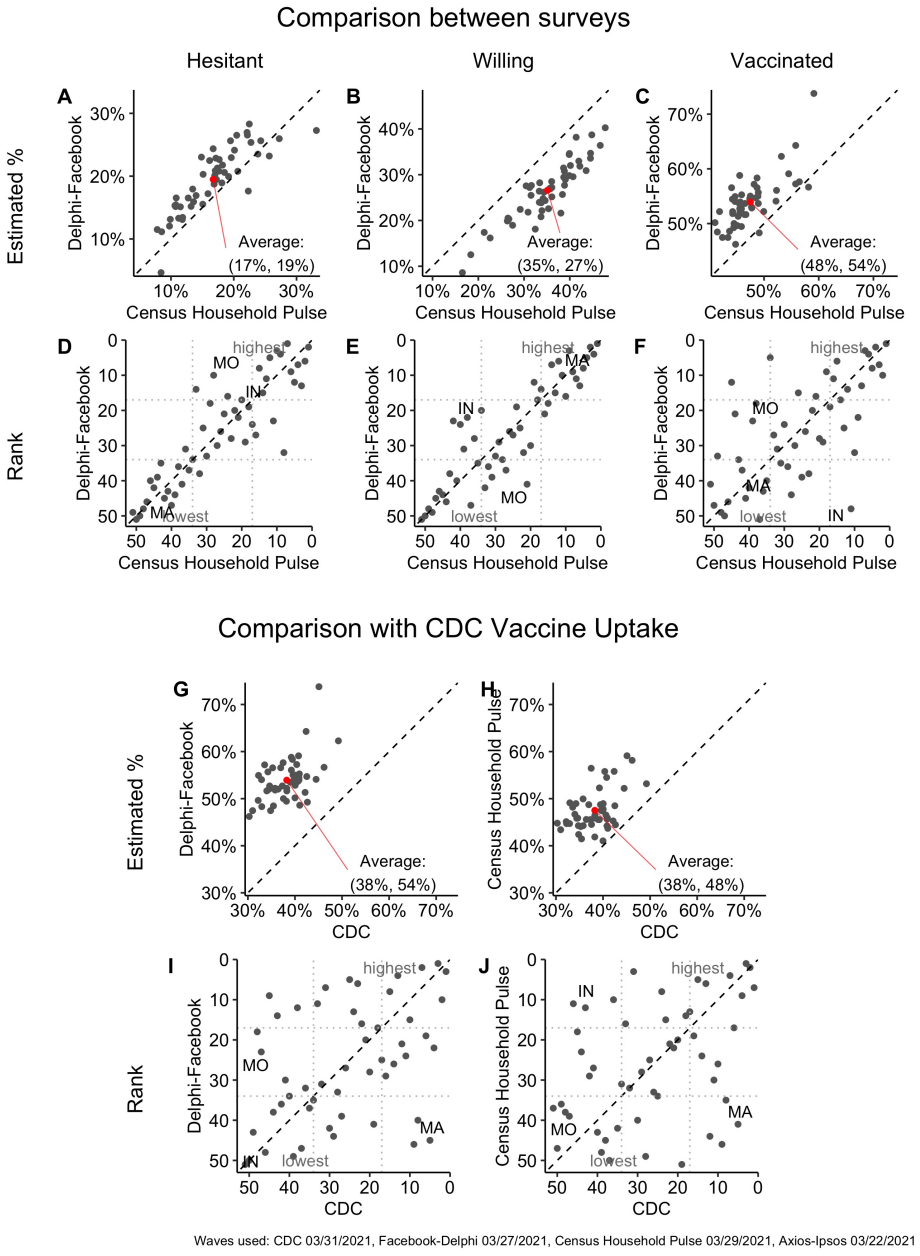


Figure 2.3: Comparisons of state-level vaccine uptake, hesitancy, and willingness across surveys and the CDC: March 2021 Comparison of state-level point estimates (A-C) and rankings (D-F) for vaccine hesitancy, willingness, and uptake from Delphi-Facebook, and Census Household Pulse. Dotted black lines show agreement and red points show the average of 50 states. Panels G-J compare state-level point estimates and rankings for the same survey waves to CDC benchmark estimates from March 31, 2021. The Delphi-Facebook data is from the week ending March 27, 2021 and the Census Household Pulse is the wave ending March 29, 2021.

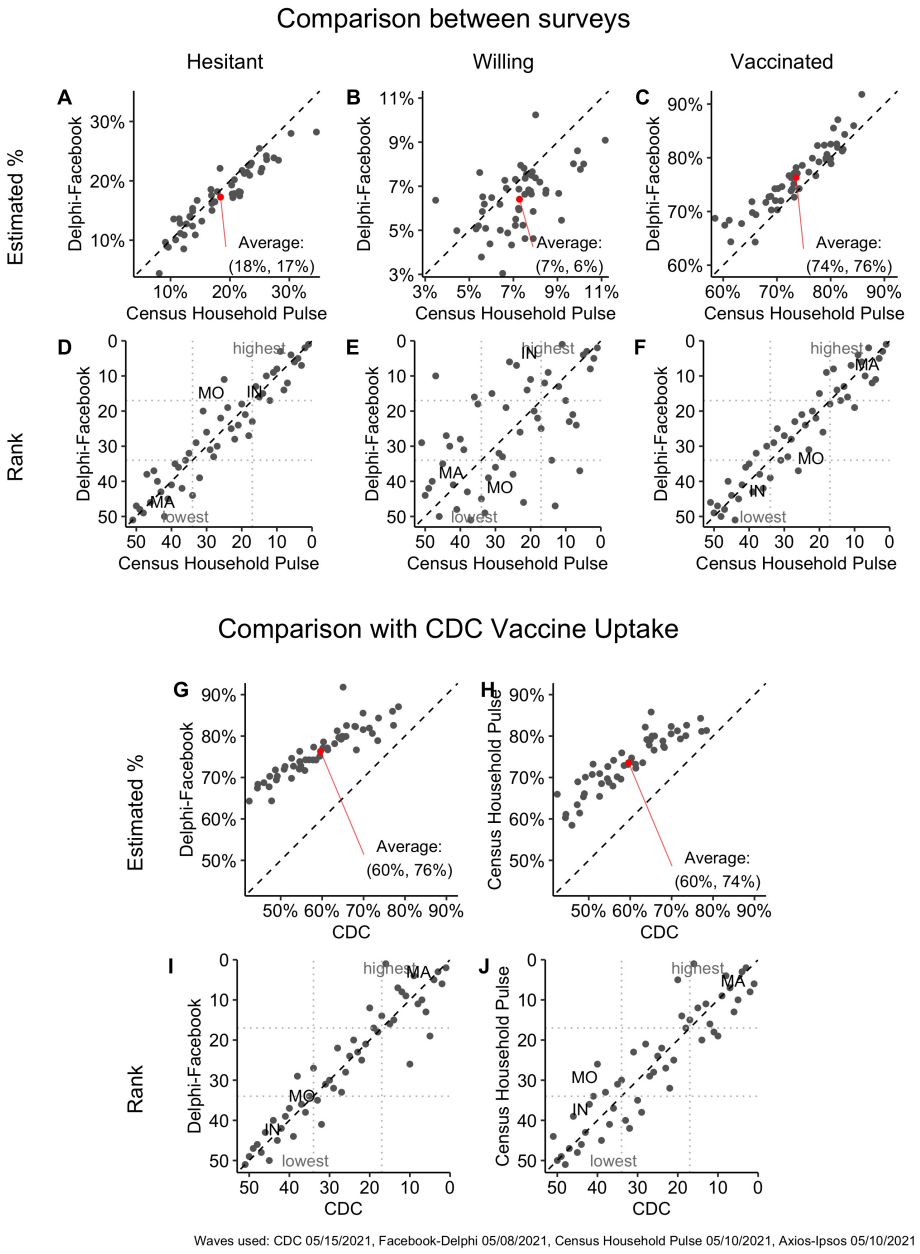


Figure 2.4: Comparisons of state-level vaccine uptake, hesitancy, and willingness across surveys and the CDC: May 2021. Comparison of state-level point estimates (A-C) and rankings (D-F) for vaccine hesitancy, willingness, and uptake from Delphi-Facebook, and Census' Household Pulse. Dotted black lines show agreement and red points show the average of 50 states. Panels G-J compare state-level point estimates and rankings for the same survey waves to CDC benchmark estimates from May 15, 2021. The Delphi-Facebook data is from the week ending May 8, 2021 and the Census Household Pulse is the wave ending May 10, 2021.

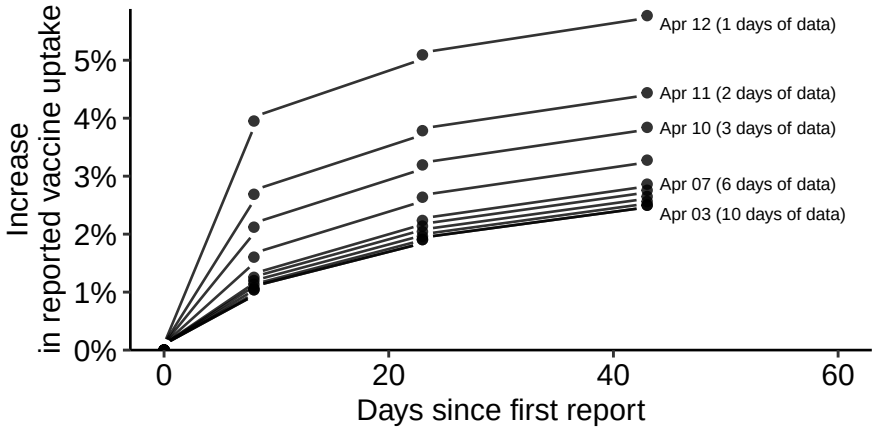


Figure 2.5: Retroactive adjustment of CDC vaccine uptake figures for April 3-12, 2021, over the 45 days from April 12. Increase is shown as a percentage of the vaccine uptake reported on April 12. Most of the retroactive increases in reported estimates appear to occur in the first 10 days after an estimate is first reported. By about 40 days after the initial estimates for a particular day are reported, the upward adjustment plateaus at around 5-6% of the initial estimate. We use this analysis to guide the choice of 5% and 10% error in the CDC benchmark for our robustness checks.

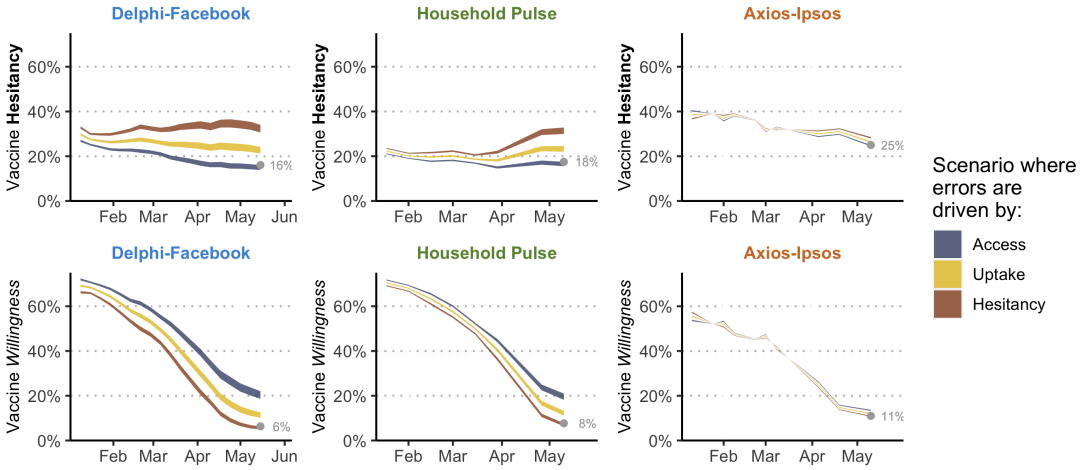


Figure 2.6: Revised estimates of hesitancy and willingness after accounting for survey errors for vaccination uptake. The gray point shows the reported value at the last point of the time series. Each line shows a different scenario for what might be driving the error in uptake estimate, derived using hypothetical *ddc* values for willingness and hesitancy based on the observed *ddc* value for uptake. *Access* scenario: willingness suffers from at least as much, if not more, bias than uptake. *Hesitancy* scenario: hesitancy suffers from at least as much, if not more, bias than uptake. *Uptake* scenario: the error is split roughly equally between hesitancy and willingness. See Supplementary Information 2.16 for more details.

	Axios-Ipsos	Census Household Pulse	Delphi-Facebook
Purpose	Measure national attitudes toward COVID-19	Sub-national social and economic impact of COVID-19	Fine-grained COVID-19 symptom surveillance
Target Pop.	18+ US general pop	18+ US general pop	18+ US general pop
Length of wave	4 days, conducted weekly	2 weeks	Daily cross-section samples, reported weekly
Average participation rate among invitees	50%	6-8%	1%
Sampling design	Inverse response propensity sampling	Systematic sample of households, adjusted for a projected response rates	Unequal-probability stratified random samples
Hesitancy / Willingness question	"How likely, if at all, are you to get the first generation COVID-19 vaccine, as soon as it's available"	"Once a vaccine preventing COVID-19 is available to you, would you..."	"If a vaccine to prevent COVID-19 were offered to you today, would you choose to get vaccinated?"
Vaccine hesitancy responses	"Not very / at all likely"	"Definitely/Probably NOT get a vaccine" or "Unsure"	"No, definitely/probably not"
Languages	English and Spanish	English and Spanish	English, Spanish, Brazilian Portuguese, Vietnamese, French, and Chinese
Report MoE or design effect	Both	Report standard errors for estimates from replicate weights	Report standard errors for estimates (does not include variance from weighting)
Sources for demographic benchmarks	2019 CPS March Supplement, party ID from recent ABC/WaPo polls	2018 ACS, 1-year estimates	2018 CPS March Supplement

Table 2.3: Methodologies of Axios-Ipsos, Census Household Pulse, and Delphi-Facebook studies. Supplements information in Table 2.1.

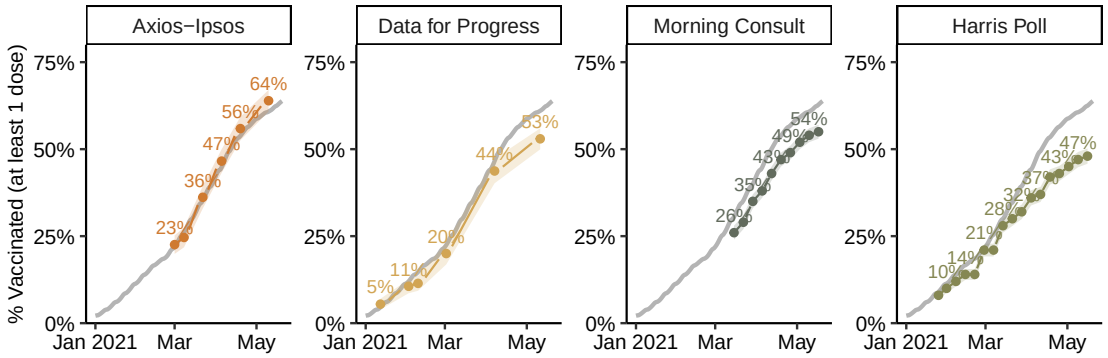


Figure 2.7: Vaccination Rates compared with CDC benchmark for four online polls. Ribbons indicate traditional 95 percent confidence intervals which are twice the standard error reported by the poll. Data for Progress asks "As of today, have you been vaccinated for Covid-19?"; Morning Consult asks "Have you gotten the vaccine, or not?"; Harris Poll asks "Which of the following best describes your mindset when it comes to getting the COVID-19 vaccine when it becomes available to you?". See the Supplementary Information 2.15.3 for more details on each survey and discussion of differences. Gray line is the CDC benchmark.

	Vaccinated		Hesitant	
	Raw	Weighted	Weighted	Sample size
Axios-Ipsos Survey				
only Offline Panelists	19%	13%	64%	21
only Online Panelists	43	37	30	974
with Ipsos Weights	42	36	30	995
with Delphi-implied Weights	42	37	29	995
Delphi-Facebook Survey				
with Delphi Weights	42%	46%	37%	249,954

Table 2.4: Contribution of offline recruitment and weighting schemes to discrepancies between surveys. A portion of each Axios-Ipsos wave is recruited from a population with no stable internet connection; Ipsos KnowledgePanel provides tablets to these respondents. In the Axios-Ipsos March 22 2020 wave, the offline panelists ($n = 21$) were 24 percentage points less likely to be vaccinated than online panelists ($n = 974$). Weighting the same Axios-Ipsos data ($n = 995$) to the age and gender target distribution implied by Delphi-Facebook's weights make the vaccination estimates higher by 1 percentage point. However, this number is still lower than Delphi-Facebook's (responses from March 14–20 2020, $n = 249,954$) own estimate of 46%. During this time period, the CDC benchmark vaccination rate was 35.2%. This suggests that the recruitment of offline respondents and different weighting schemes each explains only a small portion of the discrepancy between the two data sources.

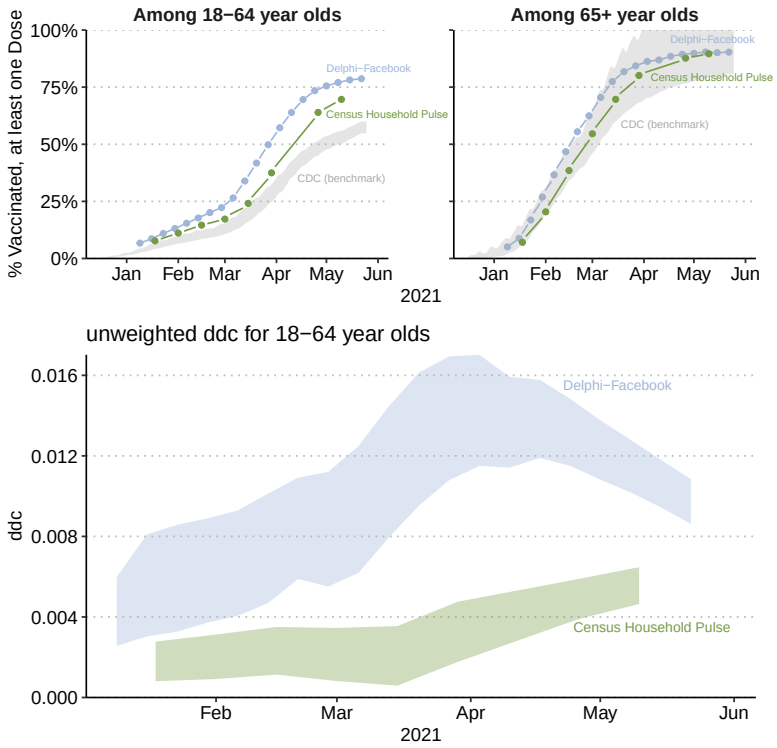


Figure 2.8: Survey error by Age Group (18-64 year-olds, and those 65 and over). a. Estimates of vaccine uptake from Delphi-Facebook (blue) and Census Household Pulse (green) for each 18-64 year-olds (left) and those 65 or older (right). Bounds on the CDC's estimate of vaccine uptake for those groups are shown in gray. The CDC receives vaccination-by-age data only from some jurisdictions. We do know, however, the total number of vaccinations in the U.S. Therefore, we calculate the bounds allocating all the vaccine doses for which age is unknown to either 18-64 or 65+. **b.** Unweighted *ddc* for each Delphi-Facebook and Census Household Pulse calculated for the 18-64 group using the bounds on the CDC's estimates of uptake. *ddc* for 65+ is not shown due to large uncertainty in the bounded CDC estimates of uptake.

Stage s	**Population N_s**	**Sampling Process $\longrightarrow$**	**Data n_s**	**n_s/N_s**
1. Define frame	144 m hh	Subset to reachable address	116 m hh	80%
2. Decide outreach list	116 m hh	Random sample	1 m adults	1%
3. Individual behavior	1 m adults	Individual responds (or not)	75,000 adults	7%
Final	$\sim$ 255 m adults		75,000 adults	0.03%

Table 2.5: Example of multi-stage population selection. The *Law of Large Populations* described in Methods section "Population size in multi-stage sampling" shows that the population size at the sampling stage where simple random sampling breaks down will dominate the error. This table explains these stages with a concrete example, using the Census Household Pulse. Population and sample sizes for three stages (stage number denoted $s \in \{1, 2, 3\}$) of sampling of the Census Household Pulse survey data collection process. Approximate sample sizes based on the March 24, 2021 wave. "m" stands for millions and "hh" stands for household. The final row compares the total adult population in the US (255 million adults, made up of 144 million households) to the sample size in one wave of the household pulse. For the purpose of illustration, we have ignored the impact of unequal sampling probabilities on the sample sizes at each stage.

2.13 Appendix: Additional Details about Data Sources

2.13.1 Total Population

The CDC vaccination data includes vaccines administered in Puerto Rico. As of June 9, 2021, approximately 1.6 million adults have received at least one dose, just under 1% of the national total (164,576,933). We use the CDC's reported national total that includes Puerto Rico (we do not have a reliable state-level time series of vaccine uptake), but we use a denominator that *does not* include Puerto Rico. This means that the CDC's estimate of vaccine uptake used here may be slightly *overestimating* the true proportion of the US (non-Puerto Rico) adult population that has received at least one dose by about 1%, which would make the observed *ddc* for Delphi-Facebook and Census Household Pulse and *underestimate* of the truth. However, this 1% error is well within the benchmark uncertainty scenarios presented with our results.

2.13.2 CDC Imputation and Uncertainty

On May 26, 2021, we obtained a daily time series of the number of cumulative first dose vaccinations that had been reported to the CDC, up to the day we obtained the data. The time series contained retroactive updates to those numbers, as do all of the CDC's daily updates to this series. In our analysis, we use CDC's values through May 19, 2021, excluding the 6 most recent days of data which suffer the most from reporting delays.

The nature of the retroactive updates suggests that in the May download, the vaccination reports for January through March are almost entirely complete, April's might still see some adjustments, and May's would be an underestimate.

To inform our prediction of how much the CDC benchmark would change in the future, we examined changes in vaccine uptake rates reported by the CDC over time. We downloaded versions of the CDC's cumulative vaccine uptake estimates that are updated retroactively as new reports of vaccinations are received on April 12, April 21, May 5, and May 26. This allowed us to examine how much the CDC's estimates of vaccine uptake for a particular day changed as new reports were received. Extended Data Fig. 2.5 compares the estimates of cumulative vaccine uptake for April 3-12, 2021 reported on April 12, 2021 to estimates for those same dates reported on subsequent dates. The top line shows that the cumulative vaccine uptake estimate for April 12, 2021 is, over the next month and a half,

adjusted upwards by approximately 6% of the original estimate reported on April 12, 2021. The estimate of vaccine uptake for April 11, reported on April 12, is further adjusted upward by approximately 4% over the next 45 days. There is little apparent difference in the amount by which estimates from April 3-8 are adjusted upwards after 45 days, indicating that most of the adjustment occurred in the first 4 days after the initial report. This is consistent with CDC's own analysis (US Centers for Disease Control and Prevention, 2021a). There is still some adjustment that occurs past day 5; after 45 additional days, estimates are adjusted upwards by an additional 2%.

We use these results to inform our choice of degree of benchmark imprecision we show in our results: 5% and 10%. The benchmark imprecision is incorporated into our analysis by adjusting the benchmark estimates each day up or down by 5% or 10% (i.e. multiplying the CDC's reported benchmark by 0.9, 0.95, 1.05, and 1.1). We then calculate *ddc* on each day for each benchmark imprecision scenario, as well as for the CDC reported benchmark, with the same survey estimate.

There are many caveats to this analysis of CDC retroactive updating, including that it depends on snapshots of data collected at inconsistent intervals, and that we mainly examine a particular window of time, April 3-12, so our results may not generalize to other windows of time. This is plausible for a number of reasons including changes to CDC reporting systems and procedures after the start of the mass vaccination program, or due to the fact that true underlying vaccine uptake is monotonically increasing over time. It is also plausible, if not likely, that the reporting delays are correlated with vaccine providers which are in turn correlated with the population receiving vaccines at a given time. As the underlying population receiving vaccines changes, so would the severity of reporting delays.

However, it is important to note that the national benchmark data that we use in our analysis of national uptake *has* been retroactively updated with new reports of vaccine administration are received with CDC's knowledge as of May 26, 2021. In other words, the possible uncertainty we consider are those *above and beyond* the reporting delays the CDC accounted for by that May date. Our benchmark imprecision calculations are intended only to test the robustness of our findings to plausible latent error in the benchmark data rather than to suggest that those scenarios are at all likely. To fully account for errors in the CDC benchmark would require a close collaboration with the CDC, and to have access to its historical information and methodologies on addressing issues such as never-reporting, as occurred when reporting AIDS status (Tu et al., 1993; Bouman et al., 2005).

Separately from the nationwide vaccine uptake, we use state-level survey estimates and CDC benchmarks in Extended Data Figs 2.3-2.4. We use the state-level vaccination counts that the CDC reports, as scraped by Our World In Data (Appel et al., 2021). Unlike the nationwide total, at the time of our study these state-level numbers were not retroactively updated as new reports of vaccines administered on previous days are reported to the CDC. so may measure the state-level vaccine uptake on any given day with the same sort of imprecision that affects the national benchmark. We use state level data only to motivate the inaccuracies of the state-level rank orders implied by vaccine uptake estimates from Delphi-Facebook and Census Household Pulse; hence they are not used to calculate *ddc*. After our study period, the CDC updated their state level time series to include retroactive updates(US Centers for Disease Control and Prevention, 2021c), but we do not analyze this data in this article.

2.13.3 Availability of Survey Microdata

Both Axios-Ipsos and Census Household Pulse release microdata publicly. Facebook also releases microdata to institutions that have signed Data Use Agreements. In view of the timeliness of our study, and to keep all three surveys on as equal a footing as possible, we used the aggregated results released by all three surveys rather than their microdata.

In all surveys, data collection happens over a multi-day period (or multi-week in the case of the Census Household Pulse). We calculate error for each survey wave with respect to the CDC-reported proportion of the population vaccinated up to and including the end date of each wave. Some respondents will have actually responded days (or weeks) before the date on which the estimate was released, when the true rate of vaccine uptake was lower. We use the end date instead of a mid-point as we do not have good data on how respondents are distributed over the response window. However, this means that the error we report may *underestimate* the true error in each survey, particularly those with longer fielding and reporting windows.

2.13.4 Census Household Pulse

The Census Household Pulse is administered by the Bureau of Labor Statistics (BLS); the Bureau of Transportation Statistics (BTS); the Centers for Disease Control and Prevention (CDC); Department of Defense (DOD); the Department of Housing and Urban Development (HUD); Maternal and Child Health Bureau (MCHB); the National Center for Education Statistics (NCES); the National Center

for Health Statistics (NCHS); the National Institute for Occupational Safety and Health (NIOSH); the Social Security Administration (SSA); and the USDA Economic Research Service (ERS) (`https://www.census.gov/programs-surveys/household-pulse-survey.html`, visited June 5, 2021). Each wave since August 2020 fields over a 13-day time window. All data used in this analysis is publicly available on the US Census website.

The Census Household Pulse changed the question used to gauge vaccine willingness and hesitancy beginning with wave 27 (the most recent wave used in this analysis), to add a response option for respondents who are "unsure" if they will receive a COVID vaccine when they become eligible. Approximately 6.6% of all respondents reported being "unsure" in wave 27, and were coded as "vaccine hesitant" rather than "willing."

2.13.5 Delphi-Facebook

Facebook performs inverse probability weighting on responses, but the reported standard errors do not include variance increases from weighting, and no estimates of design effects are released publicly. We are therefore grateful to the CMU team for providing us with estimated weekly design effects for all weeks through April 2021. The design effects are quite consistent across 2021 waves (Mean: 1.48, 95% CI: $1.48 - 1.49$), so we mean-impute the design effects for May waves.

2.14 Appendix: Asymptotic Properties of *ddc*

Here we lay out the formal results underlying the interpretation of our empirical decomposition of total error into *ddc*. The first section explains how individual response behavior drives $\hat{\rho}_{Y,R}$ and sampling rate $f = n/N$. The second section describes why the relevant population size N differs between surveys of the same target population when the data collection process involves multiple processes. This clarifies the key distinction with the classic probabilistic sampling framework, and how our results are consistent with the *Law of Large Populations* (Meng, 2018).

2.14.1 The Role of Individual Response Behavior

In the Methods "Asymptotic behavior of *ddc*", we considered a logit model of the propensity score to assert that the *ddc* $\hat{\rho}_{Y,R}$ will not vanish with the population size N, regardless of how large N is. Here we provide the mathematical proof of this assertion. First, recall that the probability calculation involving Y is with respect to its finite population $\{Y_i, i = 1, \ldots, N\}$, we have $\Pr(Y = 1) = \bar{Y}_N$. Therefore, when the individual response model

$$\Pr(R = 1|Y) = \frac{e^{\alpha+\beta Y}}{1 + e^{\alpha+\beta Y}}$$

is applicable to the entire finite population (e.g., a social media platform is open to everyone, at least in theory), we have that, as $N \to \infty$, the fraction of observations

$$f \to (1 - \mu)\frac{e^{\alpha}}{1 + e^{\alpha}} + \mu\frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} =: p, \tag{2.6}$$

where $\mu \in (0, 1)$ denotes the limit of $\bar{Y}_N$ as N increase to infinity. Here we assume such a limit exists, and it is not a trivial one (that is, μ stays away from 0 or 1). Consequently, $p \in (0, 1)$, i.e. it also stays away from 0 or 1, since it is a convex combination of $\frac{e^{\alpha}}{1+e^{\alpha}}$ and $\frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}}$, both of which lie in $(0, 1)$. This means that we cannot make the sample n arbitrarily large (or small), such as approaching N, or even at a particular level, because it is controlled by the value of $\{\alpha, \beta\}$, which is determined by the individual response behavior (towards the specific question underlying Y).

Second, because $\mathrm{Cov}(Y, R) = \mathrm{E}(YR) - \mathrm{E}(Y)\mathrm{E}(R) = \Pr(R = 1|Y = 1)\Pr(Y = 1) - \bar{Y}_N f$, we have

$$\hat{\rho}_{Y,R} \to \left(\frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} - \frac{e^{\alpha}}{1 + e^{\alpha}}\right)\frac{\sqrt{\mu(1 - \mu)}}{\sqrt{p(1 - p)}}. \tag{2.7}$$

This implies that for any given value of $\{\alpha, \beta\}$, $\hat{\rho}_{Y,R}$ will converge to a non-zero value ρ as long as $\beta \neq 0$, that is, as long as the propensity for response depends on Y itself. Consequently, the total error, relative to the standard error from simple random sampling (as a benchmark), denoted by Z,

$$Z =: \frac{\bar{Y}_N - \bar{Y}_N}{\sqrt{(1-f)\sigma^2/n}} = \hat{\rho}_{Y,R}\sqrt{N} \tag{2.8}$$

goes to infinity with N at the rate of $\rho\sqrt{N}$, a phenomenon that does not happen when $\beta = 0$.

2.14.2 Connection with the Heckman selection model

The goal of the Heckman selection model (Heckman, 1979) is to perform estimation in the case of non-response induced by censoring a latent variable. Specifically, let each member of the population be identified via a tuple of characteristics (Y_{1i}, Y_{2i}) which satisfy:

$$\begin{aligned} Y_{1i} &= X_{1i}\beta_1 + U_{1i} \\ Y_{2i} &= X_{2i}\beta_2 + U_{2i}, \end{aligned} \tag{2.9}$$

where the tuples of U_i are identically and independently distributed multivariate Normal noise:

$$\begin{pmatrix} U_{1i} \\ U_{2i} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right) \tag{2.10}$$

and the β_j's are regression coefficients. We seek to estimate β_1, but observe data Y_{1i} if and only if $Y_{2i} \geq 0$ (the predictors X_{ji} are observed for all members of the population, however). In our framework, the response indicator is $R_i = I(Y_{2i} \geq 0)$. The *ddc* ρ under the Heckman model (which is a theoretical model and hence this is a theoretical calculation) then is given by, using properties of the multivariate Normal,

$$\begin{aligned} \rho &= \text{Corr}(Y_{1i}, I(Y_{2i} \geq 0)) \\ &= r \cdot \frac{\phi(Z_i)}{\sqrt{\Phi(-Z_i)[1-\Phi(-Z_i)]}}, \end{aligned} \tag{2.11}$$

where $Z_i = -X_{2i}\beta_2/\sigma_2$. Hence in this case the *ddc* is a multiplier of the correlation r in (2.10), where the multiplier factor λ_i' resembles the inverse Mills ratio $\phi(Z_i)/\Phi(Z_i)$, where ϕ and Φ are respectively the PDF and CDF of the standard Normal $N(0,1)$.

Intuitively, it makes sense for ρ to be closely tied with r, since r drives the selection bias. For example, if $r = 0$, then Y_2 is independent from Y_1, and hence

the sign of Y_2 will carry no information about Y_1. Therefore, for the purpose of estimating β_1, the data information is not distorted by having the sample inclusion determined by the sign of Y_2, when $r = 0$. Hence $r = 0$ must imply $\rho = 0$, and vice versa. However, r alone is insufficient to capture the impact of the biased selection mechanism, since minimally the mean of Y_2, which impacts the Z term, would influence which portion of the data is more likely to be observed. The *ddc* ρ provides a metric to capture the overall effect.

In conclusion, the *ddc* framework is closely related to the framework for inferring the population mean under the Heckman selection model (corresponding to set $X_1 = 1$). The benefit of the Heckman selection model is that we can also estimate the selection mechanism itself from the observed data thanks to the distributional assumptions about the data generating mechanism. The downside of course is that the validity of our results will depend on the reliability of the assumptions. In contrast, *ddc* makes no distributional assumptions about the data generating process, and hence it is broadly applicable. However, there is no free lunch – we cannot estimate *ddc* without external information. Nonetheless, it is a useful metric in the presence of a ground truth or plausible set of scenarios for the outcome of interest, such as in our paper.

2.15 Appendix: Additional Data Analyses

2.15.1 Estimates of hesitancy by demographic groups

We show estimates of our main outcomes by Education, and then by Race, in Table 2.6. The estimates vary by mode, but the rank ordering of a particular outcome within a single survey is roughly similar across surveys. The same estimates from Household Pulse were already presented in Table 2.2.

Table 2.6: Levels of Vaccination, Willingness, and Hesitancy, estimated by demographic group. For each outcome, we estimate the same quantity from the three surveys. Axios-Ipsos (denoted AP). Census Household Pulse (denoted HP): wave ending March 29, 2021, n = 76,068. Delphi-Facebook (denoted FB): wave ending March 27, 2021, n = 181,949. These are the same waves as those in Table 2.2. Axios does not record a separate category for Asian Americans (they are lumped into "Other", so the values are left blank.

	% Vaccinated			% Willing			% Hesitant		
Education	AX	HP	FB	AX	HP	FB	AX	HP	FB
High School	28%	39%	40%	32%	40%	35%	40%	21%	25%
Some College	36	44	52	30	38	27	34	18	21
4-Year College	36	54	62	45	36	26	19	10	12
Post-Graduate	56	67	73	33	26	19	10	7	9

	% Vaccinated			% Willing			% Hesitant		
Race	AX	HP	FB	AX	HP	FB	AX	HP	FB
White	40%	50%	59%	29%	33%	24%	30%	17%	17%
Black	27	42	55	44	39	28	29	19	17
Hispanic	26	38	45	39	48	39	36	14	16
Asian		51	58		43	37		5	5

2.15.2 *ddc* by age / eligibility status across time

The CDC also releases vaccination rates by age groups, albeit not always in bins that overlap with the survey. For overlapping bins (seniors and non-seniors) we can calculate *ddc* specific to each group (Extended Data Fig. 2.8).

The CDC only receives vaccination data for age groups from certain jurisdictions, so is likely unrepresentative of the entire US adult population. Therefore, we calculate wide bounds for what the true proportion of each age group could be based on allocating the administered doses for which we do not have age information

either entirely to seniors or entirely to non-seniors. When this allocation implies a vaccination rate of more than 100% for that group, the remaining doses are allocated to the other age group. For example, if we know that on a particular day, X doses were administered to non-seniors, Y doses were administered to seniors, and Z doses were administered for which we have no age information, then the bounds for non-seniors are calculated as (X, X+Z) divided by the size of the non-senior US population. Similarly, the bounds for seniors were calculated as (Y, Y+Z) divided by the size of the senior US population.

These bounds do not incorporate any additional benchmark error, so may suffer from reporting delays or other systematic biases, and should be interpreted with caution. We do not show *ddc* for the 65+ age group due to the large width of the conservative bounds which led to unreliable estimates.

2.15.3 Other online polls

Clearly surveys can and do go wrong regardless of their sizes. Therefore, they key message of our analysis is *not* that "the smaller the better", but rather that (1) quality matters far more than quantity, and (2) large surveys fail more drastically than small surveys when there is non-negligible *ddc*. To highlight these points, we considered three more major online polls that ask vaccination status.

Figure 2.7 shows how the estimated vaccination rate of Axios-Ipsos, Data for Progress, Morning Consult, and Harris Poll tracks the CDC benchmark. The poll that is perhaps most similar to Axios-Ipsos and provides enough documentation of their methods and data, Data for Progress, generated similar patterns as Axios-Ipsos. Their estimates tended to underestimate the vaccination rate by May, but did not suffer from overconfidence in its incorrect estimate. Data for Progress is an online-only panel run in the online vendor Lucid.

Data for Progress collects samples by the online vendor Lucid. Each wave can last up to a week and has a sample size of about $n = 1,000$. They ask:

> "As you may know, vaccines for Covid-19 have now been approved by the Food and Drug Administration and are being offered to some individuals based on specific criteria. As of today, have you been vaccinated for Covid-19?" (1) "Yes, I have received at least one Covid-19 vaccination shot," (2) "No, I have not received a Covid-19 vaccination shot."

Data for Progress' poststratification weighting weights to national numbers of "gender, age, region, education, race, the interaction of education and race, and presidential vote ([2020 presidential vote])."

Harris Poll employs an online panel with an unspecified vendor. Their weekly COVID polls are about $n = 2,000$ per wave, covering three days. They ask:

> "Which of the following best describes your mindset when it comes to getting the COVID-19 vaccine when it becomes available to you?" (1) "I plan to go the first day I am able to", (2) "Whenever I get around to it", (3) "I will wait awhile and see", (4) "I will not get a COVID-19 vaccine", and (5) "I have already received a COVID-19 vaccine."

and the analysis here only takes the last option as an indicator for vaccine uptake.

The Harris Poll weights by a propensity score by their "propensity to be online," and additionally poststratify for "age, sex, race/ethnicity, education, region, household size, employment, and household income" to population benchmarks.

Morning Consult employs their own online panel. They report a margin of 1 percentage point and a rough sample size of $n = 30,000$ per week (which corresponds to a wave). They ask:

> "Have you gotten the vaccine, or not?" (1) "Yes", (2) "No, but I will get it in the future," (3) "No, and I am not sure if I will get it in the future," and (4) "No, and I do not plan to get it."

Morning Consult weights their survey data to "a range of demographic factors, including age, race/ethnicity, gender, educational attainment, and region. State-level results were weighted separately to be representative of age, gender, race/ethnicity, education, home ownership and population density."

YouGov is also a prominent online poll. However, YouGov, unlike the other polls discussed here, investigated how their estimates track the CDC vaccination rate (Blumenthal, 2021). Therefore, we do not compare it with the other polls here. They found that the "have you been vaccinated" wording was more accurate than starting the question with "will you be vaccinated?" and including an "already" option, which tended to underestimate the vaccination rate. Their A/B test confirmed the change in question wording caused a discrepancy of about 14 percentage points even in the same poll.

YouGov's A/B test provides some indication why Harris underestimates the vaccination rate. Note that Harris, unlike Data for Progress and our three surveys in the main text, uses the wording, "when [the vaccine] becomes available to you." This is precisely the type of question wording that would underestimate vaccination rates, per YouGov. The underestimation of Morning Consult may be separately due to its questions not specifying "at least one dose," thereby inducing a fraction of one-dose only respondents to not select "Yes." We therefore suspect the underestimation of the Harris Poll is due to the question wording rather than something systematic about online polls.

2.16 Appendix: *ddc*-based Scenario Analysis for Willingness and Hesitancy

The main quantity of interest in the surveys examined here is not uptake, but rather willingness and hesitancy to accept a vaccine when it becomes available. Our analysis of *ddc* of vaccine uptake cannot offer conclusive corrected estimates of willingness and hesitancy; however we propose *ddc*-based scenarios that suggest plausible values of willingness and hesitancy given specific hypotheses about the mechanisms driving selection bias.

2.16.1 Setting up scenarios

We adopt the following notation for the key random variables we wish to measure:

- V - did you receive a vaccine ("vaccination")?
- W - if no, will you receive a vaccine when available ("willingness")?
- $H = 1 - V - W$ - vaccine "hesitancy"

Just as we have studied the data quality issue for estimating the vaccine uptake, we can apply the same framework to both W and H. Unlike uptake, however, we do not have CDC benchmarks for willingness or hesitancy. We only know that $V + H + W = 1$, and therefore that

$$\text{Cov}(R, V) + \text{Cov}(R, H) + \text{Cov}(R, W) = 0$$

Re-expressing the covariances as correlation, and recognizing that $\text{Corr}(R, \cdot) = \rho_{R,\cdot}$, we obtain

$$\rho_{R,V} \cdot \sigma_V + \rho_{R,H} \cdot \sigma_H + \rho_{R,W} \cdot \sigma_W = 0$$

It is well-known that for a Bernoulli random variable, its variance is rather stable around 0.25 unless its mean is close to 0 or 1. For simplicity, we then adopt the approximation that $\sigma_V^2 \approx \sigma_H^2 \approx \sigma_W^2$. Consequently, we have

$$\rho_{R,V} + \rho_{R,H} + \rho_{R,W} \approx 0$$

As we have estimated *ddc* of vaccine uptake for each survey wave, we can further say that $\rho_{R,H} + \rho_{R,W} \approx -\hat{\rho}_{R,V}$. However, we have no information to suggest how $\rho_{R,V}$ is decomposed into *ddc* of hesitancy and willingness. Therefore, we

introduce a tuning parameter, λ, that allows us to control the relative weight given to each $\rho_{R,H}$ and $\rho_{R,W}$, such that

$$-\rho_{R,H} = (1-\lambda)\hat{\rho}_{R,V}, \quad -\rho_{R,W} = \lambda\hat{\rho}_{R,V}$$

The tuning parameter λ may take on values greater than 1 and less than -1, which would indicate that the *ddc* of either willingness or hesitancy is *greater* in magnitude than that of uptake, or that selection bias is more extreme than that or vaccine uptake.

2.16.2 Obtaining scenario estimates

Once we postulate a particular value of *ddc*, we can use Equation 2.3 to solve for the population quantity of interest, say $\bar{H}_N$. Specifically, given a postulated value of $\rho_{H,R_\text{w}} = r$, we can calculate $\bar{H}_N$ as follows:

$$\overline{H}_\text{w} - \overline{H}_N = \underbrace{r \cdot \sqrt{\frac{N-n_\text{w}}{n_\text{w}}}}_{c} \cdot \sqrt{\overline{H}_N(1-\overline{H}_N)}. \tag{2.12}$$

Squaring both sides and rearranging, we obtain:

$$(c^2+1)\bar{H}_N^2 - (2\bar{H}_\text{w} + c^2)\bar{H}_N + \bar{H}_\text{w}^2 = 0, \tag{2.13}$$

which can be solved for $\bar{H}_N$. The two roots of the quadratic equation, which we will denote by $\{h_1, h_2\}$ with $h_1 < h_2$, corresponding $\rho_{H,R_\text{w}} = r$ and $\rho_{H,R_\text{w}} = -r$. Since we know the sign of r, there will be no ambiguity on which root to take.

We note that, by setting $z = r\sqrt{N}$ and rearranging (Equation 2.12), we have

$$\frac{\overline{H}_\text{w} - \overline{H}_N}{\sqrt{\frac{1-f}{n} \cdot \overline{H}_N(1-\overline{H}_N)}} = z, \tag{2.14}$$

where $f = n_\text{w}/N$. One may recognize that is the quantity for constructing the classical Wilson score confidence interval for a binomial proportion (Brown et al., 2001), but with the finite-population correction factor $(1-f)$. This connection illuminates the meaning of the particular value of *ddc* (ρ_{H,R_w}) in this context: the quantity z, which directly depends on *ddc*, is the corresponding *quantile* used in the Wilson interval. In other words, z is the multiplier or yardstick of the benchmark error (provided by simple random sampling) to measure the error in the estimator $\overline{H}_\text{w}$. The fact that it grows with $\sqrt{N}$, when ρ_{H,R_w} does not vanish with $1/\sqrt{N}$, is precisely the explanation from the *ddc* framework.

2.16.3 Scenario estimates

We focus on three scenarios defined by ranges of λ that correspond to three mechanisms:

This allocation scheme allows us to pose scenarios implied by values of λ that capture three plausible mechanisms driving bias. First, if hesitant (H) and willing (W) individuals are equally under-represented ($\lambda \approx 0.5$), leading to over-representation of uptake, correcting for data quality implies that both Willingness and Hesitancy are higher than what surveys report (Extended Data Fig. 2.6, yellow bands). We label this the *uptake* scenario because, among the three components, uptake has the largest absolute *ddc*. Alternatively, the under-representation of the *hesitant* population could be the largest source of bias, possibly due to under-representation of people with low institutional trust who may be less likely to respond to surveys and more likely to be hesitant. This implies $\lambda \approx 0$ and is shown in the red bands. The last scenario addresses issues of *access*, where under-representation of people who are willing but not yet vaccinated is the largest source of bias, perhaps due to correlation between barriers to accessing both vaccines and online surveys (e.g., lack of internet access). This implies $\lambda \approx 1$ and upwardly corrects willingness, but does not change hesitancy.

In particular, the values used to generate the bands shown in Extended Data Fig 2.6 use the following values of lambda:

- *Access* (blue bands): $\lambda \in [1, 1.2]$, and thus $\rho_W \in [-1.2\rho_V, -\rho_V]$ and $\rho_H \in [0, 0.2\rho_V]$.
- *Hesitancy* (red bands): $\lambda \in [-1.2, -1]$, and thus $\rho_H \in [-1.2\rho_V, -\rho_V]$ and $\rho_W \in [0, 0.2\rho_V]$.
- *Uptake* (yellow bands): $\lambda \in [0.4, 0.6]$ $\rho_H \in [-0.6\rho_V, -0.4\rho_V]$ and $\rho_W \in [-0.6\rho_V, -0.4\rho_V]$.

For each of the scenarios we estimate, adjustments with $\rho_{R,V}$ (*ddc* of vaccination) by each survey puts the three survey's estimates of Hesitancy and Willingness in agreement. Because the width of each band is proportional to each survey's estimated $\rho_{R,V}$ by a constant λ, it makes sense that Delphi-Facebook has the widest band and Axios-Ipsos has the narrowest band.

The *hesitancy* scenario suggests that the actual rate of hesitancy is about 31-33% in the most recent waves of Delphi-Facebook and Census Household Pulse, almost double that of original estimates. In the *uptake* scenario, both hesitancy and

willingness are about 5 percentage points higher than each survey's original estimates. The *access* scenario suggests that willingness is as high as 21%, i.e. that a fifth of the US population still faced significant barriers to accessing vaccines as of late May.

Axios-Ipsos scenarios differ from those of the other two surveys due to its small *ddc*, and different question wording. The question that Axios-Ipsos uses to gauge vaccine hesitancy is worded differently from the questions used in Census Household Pulse and Delphi-Facebook. The question asks about likelihood of receiving a "first generation" COVID-19 vaccine, which may increase levels of hesitancy among respondents if they believe the survey is asking about an experimental, rather than a thoroughly tested, vaccine. We do see that Axios-Ipsos has markedly higher baseline levels of hesitancy than either Census Household Pulse or Delphi-Facebook. While this is likely driven in part by the lower estimated rates of vaccine uptake, it is also likely due in part to question wording. Therefore, we exclude Axios-Ipsos from our scenarios of vaccine hesitancy and willingness.

The *ddc* of Axios-Ipsos is small, its estimates of hesitancy are affected less by these scenarios. Furthermore, the implied level of Hesitancy estimates for Axios-Ipsos is higher than that of the other two polls by 5-10 percentage points in the Access scenario. In fact Axios-Ipsos' *original* estimates of Hesitancy are higher than the other polls, above and beyond demographic composition differences (Table 2.6). This is likely to the wording of the inclusion of "first generation vaccine" in Axios-Ipsos' vaccine hesitancy question (Methods section Additional survey methodology). Because such wording differences may confound the interpretation of the scenarios (*ex ante*), we do not present Axios-Ipsos' results in the same figure as the other two surveys in the main text. To be clear, the vaccination is measured in a different question than Hesitancy (Table 2.1) and does not affect our presentation of vaccination-related outcomes in earlier parts of the article.

This analysis alone cannot determine which scenario is most likely, and scenarios should be validated with other studies. However, we hope that these substantive, mechanism-driven scenarios are useful for policymakers who may need to choose whether to devote scarce resources to the Willing or Hesitant populations. Extended Data Fig. 2.6 also shows that when positing these scenarios through a *ddc* framework, the estimates from Delphi-Facebook and Census Household Pulse disagree to a lesser extent than in the reported estimates (Extended Data Fig. 2.3 and 2.4).

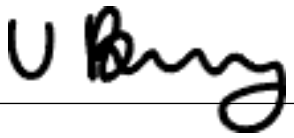
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Unrepresentative big surveys significantly overestimated US vaccine uptake
Publication Status	☒ **Published** ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X. L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, *600*(7890), 695-700.

Student Confirmation

Student Name:	Valerie C Bradley		
Contribution to the Paper	Jointly developed the idea with S.F., performed statistical analysis, wrote the manuscript.		
Signature 		Date	Jan 5, 2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: **Prof Dino Sejdinovic**		
Supervisor comments		
Signature 	Date	7 January 2024

This completed form should be included in the thesis, at the end of the relevant chapter.

3

Active Learning Sampling Design (ALSD)

Abstract

In the wake of a string of public opinion polling errors in US pre-election polling, public opinion survey research is in the midst of a transformation. Survey research is grappling with declining response rates on traditional modes and an increasingly fractured contactability landscape. To deal with these changes, the industry has turned to new modes of outreach, and more complex analysis methods like multilevel regression and post-stratification (MRP). However, these changes each tackle a single stage of the survey process (one of design, fieldwork, analysis). Therefore, we propose a new integrated framework for survey design that borrows from Bayesian optimization and active learning methods – Active Learning Sampling Design (ALSD). In ALSD, a sample is not static, but rather responsive to observed response patterns and analysis needs. The ALSD framework is intended to synthesize and provide a robust mathematical framework for many techniques already used in practical survey research. We provide the theoretical framework for ALSD and demonstrate its advantages in simulation studies.

3.1 Introduction

Over the previous decade, the field of public opinion polling has experienced a series of failures in accuracy. From the industry's failure to predict Donald Trump's victory in the 2016 US presidential election (Kennedy et al., 2018), to the largest public opinion polling miss in 40 years in the 2020 US presidential race (Clinton et al., 2020), similar misses in other countries (British Polling Council, 2016; Pennay et al., 2020a), and mixed performance in the 2022 US midterm elections. A number of factors have contributed to these misses, including political realignment in the US, declining response rates on traditional phone survey modes (Kennedy and Hartig, 2019), and the explosive growth of online nonprobability surveys. The accessibility of cheap online surveys has, in turn, contributed to a large increase in the number (and variation in quality) of pollsters releasing forecasts. Kennedy et al. (2023) find that 17% of pollsters used multiple survey modes (compared to only 2% in 2016) and a majority of firms that released polls in both 2016 and 2020 changed their methods between cycles. In this tumultuous era, there is more appetite than ever for methodological innovation in survey research.

Survey research can be (simplistically) described in four stages: design, fieldwork, analysis, and communication. At the design stage, units are selected for the study. Responses from sampled individuals are collected during fieldwork, and weighted, modeled, or otherwise analyzed at the analysis stage, and every stage, researchers seek to minimize total survey error.

In standard survey design, a variety of fieldwork and analysis needs are considered in the design stage. For example, if fieldwork is to be done in-person, the sample design may include cluster sampling (rather than simple random sampling). Or, if the goal of a survey is small area estimation, researchers may set a minimum required sample size in geographic areas of interest in order to achieve a desired level of precision at the analysis stage. However, these design choices are often static. For example, minimum sample size calculations are often based on priors from similar studies, rather than outcomes actually observed in the course of data collection. Furthermore, these traditional minimum sample size calculations are out-of-touch with increasingly popular model-based estimation methods, in particular multi-level regression and post-stratification, MRP, (Park et al., 2004, 2006).

Adaptive survey design literature proposes a more nimble approach to fieldwork, proposing methods for adapting fieldwork techniques (e.g. targeting specific individuals to receive additional follow-up attempts, or altering the mode of a follow-up attempt) based on feedback from pilot rounds of fieldwork. However,

adaptive survey design stops short of fully integrating survey analysis into sampling design (see Section 3.2.3 for a review of ASD).

Here we propose a survey design framework that allows for more iteration between the design, fieldwork, and analysis stages of survey research. We are interested in a setting in which we observe a *census* of auxiliary data for the population $\mathbf{X}_i \sim p_{\mathbf{X}}$ (although the population information is often collected from other surveys or modeled rather than a true census) and would like to design a survey to measure an outcome Y and auxiliary covariates $\mathbf{X}$ for a sample of the population $(Y_i, \mathbf{X}_i) \sim p_{Y,U,X|R=1}$ in order to estimate $\bar{Y}_N$, where R is the *response indicator*, and $R_i = 1$ for unit i if it is observed in the sample, and 0 otherwise. Active Learning Sampling Design (ALSD) borrows from the active learning and Bayesian Optimization literature to fully integrate analysis into the design of the sample used to estimate $\bar{Y}_N$. In particular, we propose a framework for dynamic sampling that customizes sampling to best suit the analysis needs of a particular survey (e.g. MRP), and adapts sampling strategies to account for observed heterogeneous response rates, thus producing more efficient samples.

The rest of the paper proceeds as follows. In Section 3.2 we provide an overview of related methods in survey literature and an introduction to active learning and Bayesian optimization, upon which our method builds. Section 3.3 outlines the ALSD framework and Section 3.4 presents results from simulation studies evaluating ALSD relative to existing sampling methods. Section 3.5 concludes.

3.2 Related Methods

3.2.1 Inverse Probability Weighting (IPW)

While survey researchers take care in the design and fieldwork phases to minimize the potential for nonresponse bias, it is typically still necessary to correct for lingering nonresponse bias in the analysis stage, often using Inverse Probability Weighting (IPW) (Horvitz and Thompson, 1952b; Deville and Särndal, 1992). IPW methods produce a set of weights for respondents such that the weighted sample distributions match specified population targets for a set of adjustment variables X, henceforth referred to as features.

Raking and post-stratification are two of the most common IPW methods, and are part of a family of calibration weighting methods. In general, calibration weights are those that are closest to the sampling design weights, subject to the constraint that weighted sample distributions must match defined population target distributions. Specific calibration weighting methods are defined by the metric used

to measure the distance between prior and final weights, and by how population targets are defined (i.e. using joint or marginal distributions). For example, raking minimizes the cross entropy between the design weights and adjusted weights subject to constraints defined by the *marginal* target distributions. In post-stratification, constraints are defined by the full joint distribution of target variables.

IPW estimators will be unbiased if the *conditional ignorability* assumption holds, which states that an outcome Y measured in a survey is conditionally independent of the response mechanism R ($R_i = 1$ if unit i is observed, 0 otherwise) given the adjustment set X, such that $Y \perp\!\!\!\perp R|X$.

IPW decreases estimator bias at the cost of increased variance. This cost of a set of weights w is measured by Kish's design effect (Kish, 1965), or *deff*:

$$\text{deff}(w) = 1 + \frac{\text{var}(w)}{\bar{w}^2} \tag{3.1}$$

The *deff* can be used to calculate the *effective sample size*, $n_w = n/\text{deff}(w)$, which indicates the size of a simple random sample with the same expected MSE as the observed weighted sample. When sample distributions match target population distributions without adjustment, $w = 1$ for all sample units, thus $\text{var}(w) = 0$, there is no variance inflation from weighting, and $n_w = n$. However, we use covariates to correct for observable nonresponse bias, weights will deviate from 1, increasing the design effect and decreasing the effective sample size.

3.2.2 Multilevel Regression and Post-stratification (MRP)

A model-based alternative to IPW for adjusting for nonresponse bias in surveys is multilevel regression and post-stratification, or MRP, (Park et al., 2004, 2006). MRP uses a multilevel, often Bayesian, model to produce estimates for fine-grained population strata with appropriate pooling and regularization to help improve estimates of sparse strata. The stratum-level estimates are then aggregated using post-stratification weighting to correct for any observed selection bias.

MRP is particularly well-suited for using survey data collected at one level of geography to estimate outcomes at a more granular geographic level. MRP is widely used across a range of application areas, including political forecasting (Gelman et al., 2016; Lauderdale et al., 2020; Gelman, 2021), public health outcomes (Zhang et al., 2014; Verity et al., 2020; Park et al., 2022), and small area estimation (Jackman et al., 2019; Chenevert et al., 2017).

3.2.3 Adaptive survey design (ASD)

Adaptive (or "responsive") survey designs tailor survey design features (e.g. mode, incentives, number of attempts) based on earlier phases of a survey in order to minimize error or cost (Groves and Heeringa, 2006), or maximize response rates (Wagner, 2008). Adaptive survey designs occasionally use Bayesian frameworks, as in Coffey (2020), or Bayesian priors for estimating survey response rates in the survey design phase (Coffey et al., 2020; West et al., 2023).

The key difference between adaptive survey design and our proposed ALSD is that adaptive survey design assumes that the sample itself is static, and only contact strategies are optimized, whereas ALSD adapts the sample itself based on the responses that have (or have not) been collected already.

3.2.4 Error decomposition and *ddc*

In motivating ALSD, it is helpful to recall the decomposition of error in a weighted estimator $\hat{Y}_w$ for a population mean $\bar{Y}_N$ from Meng (2018):

$$\bar{Y}_w - \bar{Y}_N = \underbrace{\hat{\rho}_{Y,R_w}}_{\text{data defect correlation}} \times \underbrace{\sqrt{\frac{N}{n_w} - 1}}_{\text{data sparsity}} \times \underbrace{\sigma_Y}_{\text{population heterogeneity}}$$

where $\hat{\rho}_{Y,R_w}$ measures the correlation between an outcome Y and the weighted response mechanism $R_{w,i} = w_i R_i$, where w here is a set of survey weights used to correct for nonresponse bias. If the conditional ignorability assumption holds, then $Y \perp\!\!\!\perp R_w$ (where w captures sufficient information from the adjustment set $\mathbb{X}$), and $\mathbb{E}[\hat{\rho}_{Y,R_w}] = 0$. Meng (2018) shows that probability samples control the MSE of the estimator $\bar{Y}$ by ensuring that $\mathbb{E}[\hat{\rho}_{Y,R}] = N^{-1/2}$, however when the probabilistic nature of a sample is violated (e.g. due to nonresponse), any control over selection bias is lost, and even small *ddc* can produce severe error. Thus, the *ddc* captures the sign and degree of selection bias.

The error of a sample estimator can be further decomposed by stage of the data collection process (Bradley et al., 2021). Say that $R_w = R_1 \times R_2 \times R_3$ where stages 1-3 are defined as follows:

	Stage	Population size	Sample size	$R^{(s)}$	$R^{(s)}$ is probabilistic?	$\bar{Y}$
1	Design	N	n_1	$R^{(1)}$	Yes	$\bar{Y}_{n_1}$
3	Fieldwork	n_1	n_2	$R^{(2)}$	No	$\bar{Y}_{n_2}$
3	Analysis	n_2	n_3	$R^{(3)}$	No	$\bar{Y}_w$

Table 3.1: Stages of survey data collection that contribute to overall sample error.

Thus,

$$\bar{Y}_w - \bar{Y}_N = [\bar{Y}_w - \bar{Y}_{n_2}] + [\bar{Y}_{n_2} - \bar{Y}_{n_1}] + [\bar{Y}_{n_1} - \bar{Y}_N] \tag{3.2}$$

$$= \hat{\rho}^{(3|2)}_{R^{(3)},Y} \sqrt{\frac{n_2 - n_3}{n_3}} \hat{\sigma}_{n_2} + \hat{\rho}^{(2|1)}_{R^{(2)},Y} \sqrt{\frac{n_1 - n_2}{n_2}} \sigma_{n_1} + \hat{\rho}_{R^{(1)},Y} \sqrt{\frac{N - n_1}{n_1}} \sigma_N \tag{3.3}$$

The ultimate goal of survey design, fieldwork, and analysis is to minimize the total error of $\bar{Y}_w$, via minimizing the *ddc* and data sparsity at each stage. For example, in the design phase, stratified random sampling may be used to control the sample variance and thus the effective sample size (minimize *data sparsity*), and researchers strive to ensure that the sampling frame does not suffer from coverage bias (minimize *ddc*). During fielding, we attempt to maximize the number of respondents (minimize *data sparsity*), and may use quotas to minimize nonresponse bias (minimize *ddc*). At the analysis stage, weighting or modeling is used to correct for any observed bias (minimize *ddc*) without paying too high a cost in variance inflation (minimize *data sparsity*).

3.2.5 Quota and response rate sampling

Quota sampling methods are used to force survey responses to appear representative of the population of interest with respect to a set of pre-specified, observable characteristics (Berinsky, 2006). With quota sampling, the target population is divided into strata based on key demographic characteristics believed to explain the response mechanism R and outcome Y, and target numbers of respondents set for each stratum. Then, potential respondents are selected by the researcher (not necessarily at random), and responses collected until the target number in each stratum has been reached. However this approach only ensures that sample distributions will match population distributions for a limited set of covariates and does not guarantee true *statistical* representativeness driven by a random selection mechanism, which ensures that sample distributions are unbiased with respect to every possible feature. Statistical representativeness is a nuanced concept which we will not discuss here, but see Kruskal and Mosteller (1979a,b, 1980) for more in-depth discussions.

Response rate sampling attempts to correct for this by designing a stratified random sample such that the probability of selection in each stratum is inversely proportional to the predicted response rate in that stratum (Hartman, 2014). Response rates used for designing the sample are based on actual response rates observed in previous similar surveys.

In the context of Meng's error decomposition, sampling inversely proportionally to predicted response rates can be thought of as designing a sample to have a *ddc* that will counterbalance the anticipated *ddc* at the response stage. In terms of the stages listed in Table 3.1, response rate sampling aims to select a sample such that $-\hat{\rho}^{(2|1)}_{R^{(2)},Y} = \hat{\rho}^{(3|2)}_{R^{(3)},Y}$. It is worth noting that in order for the error in stage 2 to perfectly cancel out the error in stage 3, it would actually be necessary to select a sample that accounted for the relative sampling rates at stage 2 and 3, such that $-c \times \hat{\rho}^{(2|1)}_{R^{(2)},Y} = \hat{\rho}^{(3|2)}_{R^{(3)},Y}$, where $c = \sqrt{\frac{n_2-n_3}{n_3}}/\sqrt{\frac{n_1-n_2}{n_2}}$.

The benefit of response rate sampling over simply enforcing respondent quotas is the ability to adjust for a larger set of features $\mathbf{X}$. In practice, it can be difficult, if not prohibitively inefficient, to enforce joint quotas of 3 or more features. However, with stratified response rate sampling, it is possible to account for more features in both the stratification and response rate modeling steps, giving more flexibility in selecting $\mathbf{X}$ to ensure that $Y \perp\!\!\!\perp R|\mathbf{X}$, as is required to satisfy the conditional ignorability assumption and recover from selection bias (Little and Rubin, 2019).

3.2.6 Bayesian Optimization and Active Learning

Bayesian optimization (BO) and active learning (AL) are machine learning techniques used to improve model performance and reduce the amount of data needed for training (Garnett, 2023; Settles, 2009). Bayesian optimization, and adaptive decision analysis more broadly, has been used in a wide variety of applications, including experimental design and machine learning, and has been identified as one of the most important statistical ideas of the last 50 years (Gelman and Vehtari, 2021).

BO is commonly used for optimizing expensive black-box functions. It is broadly applied to model hyperparameters search (Turner et al., 2021). BO uses a probabilistic surrogate model of the objective function to iteratively select the most promising point to evaluate next. This choice is made on the principle of exploration-exploitation trade-off by selecting the next point as the maximum of an acquisition function. A convenient acquisition function is for example, $\mu(x) + \gamma\sigma(x)$, where $\mu(x)$ and $\sigma(x)$ are the estimated mean and uncertainty measures computed based on the data observed so far. The surrogate model is updated after each evaluation, incorporating the new information to refine the estimate $\mu(x)$ of the true objective function. This process continues until a satisfactory solution is found.

AL is a technique used to select the most informative data points for training a model. It involves iterative querying an oracle (a human or an expensive black-box function) to label the most informative data points that would help the model to generalize better. The model is trained on this labeled data and the process is

repeated until the model performance reaches a desired level or until the available budget for labeling is exhausted. AL is particularly useful in situations where labeled data is scarce and expensive to obtain.

3.3 ALSD

In this section we introduce ALSD. First, we establish notation we will use, then present the general framework of ALSD, and finally discuss tailoring ALSD to fit specific analysis needs, for example whether the survey aims to maximize efficiency for a range of outcomes, or focus on a specific outcome.

3.3.1 Notation

We assume that we have access to a sampling frame of units in the population $i = 1, \ldots, N$, and observe some set of features $\mathbf{X}_i$ for all population units. We collect data over a series of survey waves $j = 1, \ldots, J$, and select new samples to invite to participate in each wave. We use ω to denote the current wave.

In each wave j, we select a sample from the population ($S_i^j = 1$ if unit i is sampled in wave j, 0 otherwise), and observe responses from some subset of the sample ($R_i^j = 1$ if unit i responds in wave j, 0 otherwise; if $S_i^j = 0$ then $R_i^j = 0$ by definition). The sample for wave ω is designed to achieve a *target* sample size of t_ω responses (we will provide more detail on exactly how this is determined below), while n_ω denotes the number of *actual* responses in wave ω.

If a unit has already responded, it will not be sampled again. Hence, a unit may only *respond* to a survey once ($\sum_j R_i^j \in \{0, 1\}$), but may be *sampled* in more than one wave if they have failed to respond to prior survey invitations ($\sum_j S_i^j \in \{0, \ldots, J\}$). Each unit has some true latent probability of responding to our survey, conditional upon being invited to participate, $\pi_i = \mathbb{P}(R_i^j = 1 | S_i^j = 1)$, and we assume that this probability remains approximately constant across waves (i.e. does not depend on j). At wave ω, we denote an indicator of whether unit has responded up to and including wave ω by $Q_i^\omega = \sum_{j=1}^{\omega} R_i^j$.

We assume that we have a fixed total budget B to spend on data collection, and that cost is measured *per completed response*. For simplicity, we further assume that the cost is constant across the population and across waves $c_i^j = c = 1 \; \forall i \in \{1 \ldots N\}, j \in \{1, \ldots J\}$. In each wave, after responses are observed, the cost of that wave is calculated as $c^j = \sum_{i=1}^{N} R_i^j$. If we are under budget with $\sum_{j=1}^{\omega} c^j < B$, then we proceed to wave $j = \omega + 1$, otherwise we end data collection.

We divide the sampling frame into strata $\mathcal{A}_h$, for $h = 1, \ldots, H$ based on the joint distribution of features $\mathbf{X}$. For example, if $\mathbf{X}$ includes age and gender, then each strata contains units with a particular combination of age and gender. As age is continuous, it would generally be discretized into buckets prior to stratification such that strata take on values such as "women x age under 30" or "men over 65", etc.. In general, the goal of stratification is to divide the population such that the intra-stratum variance of Y_i and R_i is low, but the inter-stratum variance of stratum means $\bar{Y}_h$ and $\bar{R}_h$ is high. See Lohr (2010) for more practical guidance on stratification for sampling. We denote the population total in stratum $\mathcal{A}_h$ by N_h and the cumulative number of observed units within it by n_h. Un-responded units up to and including wave ω in stratum $\mathcal{A}_h$ will be denoted $\mathcal{A}_h^\omega$, i.e. $\mathcal{A}_h^\omega = \{i \in \mathcal{A}_h : Q_i^\omega = 0\}$.

3.3.2 ALSD Overview

At a high level, ALSD follows the following steps:

1. **Initialization**. Set the overall budget B, initial target sample size t_1, and target sample size for waves $j > 1$. Specify analysis goals and methods. Finally, select an initial sample.

2. **Field survey**. Invite sampled units to participate in the survey and observe their responses. Calculate cost of the wave ω as $c^\omega = c \sum_i R_i^\omega$, and remaining budget $B^{\omega+1} = B^\omega - c^\omega$.

3. **Model response propensity**. Fit model for the response propensity as a function of the observed features, i.e. $\hat{\pi}_i = \hat{\mathbb{P}}(R_i^\omega = 1 | S_i^\omega = 1) = f(\mathbf{X}_i)$. Estimate the average response propensity of un-responded units in stratum h as $\hat{\pi}_h = \frac{1}{N_h - n_h} \sum_{i \in \mathcal{A}_h^\omega} \hat{\pi}_i$.

4. **Set the probability of selection for the next wave.** We set $\mathbb{P}(S_i^{\omega+1} = 1) = p_h^{\omega+1}$, for $i \in \mathcal{A}_h^\omega$, i.e. selection probabilities are determined at the level of stratum. Note that $\mathbb{P}(S_i^{\omega+1} = 1) = 0$ if the unit has responded before, i.e. $Q_i^\omega = 1$.

5. **Select sample for next wave**. Randomly sample from the set of units that has not yet responded using stratified random sampling where a unit i in stratum h has probability p_h of being selected.

6. **Repeat steps 2-5 until the budget is exhausted**.

3.3.3 Initialization

The overall budget B should be determined by practical constraints, and should be divided between an initial sample and subsequent waves. There is a trade-off between initial target sample size (t^1) and subsequent sample sizes. A larger t^1 will produce more precise estimates of $\hat{\pi}_i$ and $\hat{Y}$, however leaves less room to adapt future samples.

Once B and t^j are set, the initial sample may be selected by simple random sampling (SRS), or perhaps stratified random sampling where strata are defined using some subset of available features $\mathbf{X}$. For simplicity, we assume that the probability of selection in wave 1, $\mathbb{P}(S_i^1 = 1)$ is the same for all units, but that we have some rough estimate of the average response rate in the population, $\bar{\pi}_N$, such that $\mathbb{P}(S_i^1 = 1) = t^1/\bar{\pi}_N$, where t^1 is the *target* sample size for wave 1.

3.3.4 Modeling $\mathbb{P}(R_i^j = 1|S_i^j = 1)$

Hartman (2014) suggests using a tree-based method for modeling response propensity $\mathbb{P}(R_i^j = 1|S_i^j = 1)$ in order to handle interactions and mimic the structure of stratified random sampling, however any probabilistic classification method will do. In public opinion polling, response propensity across a sample is often quite low, e.g. <10% (Kennedy and Hartig, 2019), which can be difficult to model precisely. We therefore recommend oversampling units that have responded relative to those that have not, and then adjusting the average of the fitted model to match the true observed mean. This may not be necessary in government surveys with much higher average response propensities.

We use the cumulative set of samples and responses from waves $1, \ldots, \omega$ to model $\mathbb{P}(R_i^{\omega+1} = 1|S_i^{\omega+1} = 1)$. Concretely, $\mathbb{P}(R_i^j = 1|S_i^j = 1) \sim f(\mathbf{X}_i)$, such that if unit i was sampled in wave 2 but *did not* respond, and sampled in wave 4 and *did* respond, then they are included in the modeling set twice with $R_i^2 = 0$ and $R_i^4 = 1$.

Note that if respondents are over-sampled in order to more precisely model $\mathbb{P}(R_i^j = 1|S_i^j = 1)$, the fitted values will have to be calibrated such that they represent (approximately) true probabilities of response. See Rosenman et al. (2023) for a discussion of predictive model calibration.

3.3.5 Specifying $p_h^{\omega+1}$

In the typical active learning setting, the goal of data acquisition is to minimize the total predictive uncertainty of a modeled outcome by labeling points with the most predictive uncertainty. Survey sampling presents challenges that are less common in active learning: 1) extremely low and heterogeneous response rates,

2) high costs (logistic and financial) of performing additional rounds of collection and 3) many outcomes of interest. Unlike in the standard active learning setting, observations must be batch labeled instead of one at a time and batch selection must account for heterogeneous nonresponse. Furthermore, because surveys are so financially and logistically burdensome to implement, there are typically a wide range of outcomes for which population inferences will be generated. Therefore, in most cases, we cannot simply seek to minimize the uncertainty for a single outcome, but instead must attempt to minimize uncertainty for all possible outcomes. In order to do this, rather than seeking to minimize the variance of a single outcome, we will aim to minimize the variance inflation from weighting (measured by the *deff*) that impacts every outcome measured in a survey.

Therefore, there are two key components of our approach to setting the probability of selection in each wave of ALSD: 1) response rate sampling and 2) setting the target number of respondents per stratum $t_h^{\omega+1}$ to minimize the design effect. Response rate sampling allows us to have a clear understanding of the distribution of respondents expected from a particular sample before fielding and without having to resort to discarding respondents from setting quotas. Second, selecting a sample that minimizes the expected design effect minimizes the variance inflation from weighting, and maximizes the effective sample size, thus producing an efficient sample in an outcome-agnostic way.

We start by defining the probability of selection in wave $\omega + 1$ as

$$p_h^{\omega+1} = t_h^{\omega+1} \times \underbrace{\frac{1}{(N_h - n_h)}}_{\substack{\text{sampleable units} \\ \text{in stratum } h}} \times \underbrace{\frac{1}{\hat{\pi}_h}}_{\substack{\text{response rate} \\ \text{of sampleable units}}} \tag{3.4}$$

where $\sum_h t_h^{\omega+1} = t^{\omega+1}$ and $(N_h - n_h)\hat{\pi}_h$ is the maximum number of responses that could be expected from h if all remaining units are sampled. Thus, p_h is the fraction of total expected responses from stratum h that we would like to observe.

The rest of the section describes each component, response rate sampling and deff minimization, in greater detail.

Response rate sampling

We tackle the problem of large heterogeneous nonresponse by selecting a batch of samples each wave of ALSD using stratified random sampling where the probability of selection is inversely proportional to predicted response rates, as introduced in Section 3.2.5.

Response rate sampling allows us to design a sample with a precise control over the number of respondents, and their distribution across strata. Consider the expected value of respondents in wave $\omega + 1$,

$$\mathbb{E}\left[\sum_i R_i^{\omega+1}\right] = \sum_i \mathbb{E}[R_i^{\omega+1}] \tag{3.5}$$

$$= \sum_i \left(\mathbb{E}[R_i^{\omega+1}|S_i^{\omega+1}=1]\mathbb{P}(S_i^{\omega+1}=1) + \mathbb{E}[R_i^{\omega+1}|S_i^{\omega+1}=0]\mathbb{P}(S_i^{\omega+1}=0)\right). \tag{3.6}$$

By definition, only units that are sampled are able to respond, hence $\mathbb{E}[R_i^{\omega+1}|S_i^{\omega+1} = 0] = 0$, and thus

$$\mathbb{E}\left[\sum_i R_i^{\omega+1}\right] = \sum_i \mathbb{E}[R_i^{\omega+1}|S_i^{\omega+1}=1]\mathbb{P}(S_i^{\omega+1}=1) \tag{3.7}$$

$$= \sum_{i:Q_i^\omega=0} \mathbb{E}[R_i^{\omega+1}|S_i^{\omega+1}=1]\mathbb{P}(S_i^{\omega+1}=1), \tag{3.8}$$

where the second line follows since the unit may only be sampled if it has not responded in a prior wave $j < \omega + 1$, i.e. $Q_i^\omega = 0$. Therefore, breaking the sum across strata

$$\mathbb{E}\left[\sum_i R_i^{\omega+1}\right] = \sum_h \sum_{i\in\mathcal{A}_h^\omega} \mathbb{P}(R_i^{\omega+1}=1|S_i^{\omega+1}=1) \times p_h^{\omega+1}. \tag{3.9}$$

Substituting in the definition of $p_h^{\omega+1}$ from Equation 3.4, $p_h^{\omega+1} = t_h^{\omega+1} \times \frac{1}{(N_h - n_h)} \times \frac{1}{\hat{\pi}_h}$:

$$\mathbb{E}\left[\sum_i R_i^{\omega+1}\right] = \sum_h t_h^{\omega+1} \times \frac{1}{\hat{\pi}_h} \times \frac{1}{(N_h - n_h)} \times \sum_{i\in\mathcal{A}_h^\omega} \mathbb{P}(R_i^{\omega+1}=1|S_i^{\omega+1}=1). \tag{3.10}$$

Note that the final term in the equation above is the sum of true response rates over all units in stratum h that have not yet responded, of which there are exactly $N_h - n_h$. Therefore, if $\hat{\pi}_h$ is the *true average response rate of un-responded units* in stratum h, i.e.

$$\hat{\pi}_h = \sum_{i\in\mathcal{A}_h^\omega} \mathbb{P}(R_i^{\omega+1}=1|S_i^{\omega+1}=1)/(N_h - n_h) \tag{3.11}$$

then the expectation simplifies to

$$\mathbb{E}\left[\sum_i R_i^{\omega+1}\right] = \sum_h t_h^{\omega+1} = t^{\omega+1} \tag{3.12}$$

where $t^{\omega+1}$ is the target number of responses in wave $\omega + 1$. We note that it suffices that the modeled response rate probabilities are accurate at the stratum level (e.g. if they are unbiased at the individual level with each stratum having a large size) in order to obtain precise control over the expected number of respondents.

Minimizing *deff*

The second component of our approach to setting $p_h^{\omega+1}$ is to minimize the design effect. Recall that when a sample does not require weighting (i.e. is observably representative), the weights are all equal to 1, and thus the weighting design effect is 1. A sample that does not require weighting will be maximally efficient in terms of the effective sample size, $n_w = n/\text{deff}(w) = n$. Therefore, in order to maximize the expected effective sample size of the cumulative sample, we should set selection probabilities for the next wave to minimize the expected design effect.

Therefore we next define the design effect of post-stratification weights in terms of $t_h^{\omega+1}$. As we are interested in stratum-level selection probabilities, we focus on the design effect of post-stratification weights, which have the added benefit of having a closed-form solution (unlike other more iterative weighting methods).

In general, post-stratification weights for units in stratum h are given by $w_h = \frac{N_h/N}{n_h/n}$. However, we wish to modify these weights to account for the responses we expect to observe in the next wave of sampling, $t_h^{\omega+1}$

$$w_h^{\omega+1} = \frac{N_h/N}{(n_h + t_h^{\omega+1})/(n + t^{\omega+1})} = \frac{N_h(n + t^{\omega+1})}{N(n_h + t_h^{\omega+1})} \tag{3.13}$$

The mean of the weights is 1:

$$\begin{aligned}
\text{mean}(w^{\omega+1}) &= \frac{\sum_h \sum_{i \in \mathcal{A}_h} w_h^{\omega+1}}{n + t^{\omega+1}} && (3.14)\\
&= \frac{\sum_h (n_h + t_h^{\omega+1}) w_h^{\omega+1}}{n + t^{\omega+1}} && (3.15)\\
&= \frac{1}{n + t^{\omega+1}} \sum_h \frac{(n_h + t_h^{\omega+1}) N_h (n + t^{\omega+1})}{N(n_h + t_h^{\omega+1})} && (3.16)\\
&= \sum_h \frac{N_h}{N} && (3.17)\\
&= 1 && (3.18)
\end{aligned}$$

Thus, the expected design effect is

$$\begin{aligned}
\text{deff}(w^{\omega+1}) &= 1 + \frac{\text{var}(w^{\omega+1})}{\text{mean}^2(w^{\omega+1})} && (3.19)\\
&= 1 + \sum_h \sum_{i \in \mathcal{A}_h} \left(w_h^{\omega+1} - 1 \right)^2 && (3.20)\\
&= 1 + \sum_h (n_h + t_h^{\omega+1}) \left(\frac{N_h(n + t^{\omega+1})}{N(n_h + t_h^{\omega+1})} - 1 \right)^2 && (3.21)
\end{aligned}$$

Simple strategy

The simplest strategy for minimizing the design effect would be to ensure that the weights are all 1, or

$$\frac{N_h}{N} = \frac{n_h + t_h^{\omega+1}}{n + t^{\omega+1}}, \tag{3.22}$$

which can be achieved if

$$t_h^{\omega+1} = \frac{N_h}{N}(n + t^{\omega+1}) - n_h. \tag{3.23}$$

The solution is intuitive in that the optimal target number of respondents from stratum h is equal to the proportion of the population that falls into stratum h, N_h/N, times the cumulative number of responses after the next wave, $n + t^{\omega+1}$, less the number of responses already collected in stratum h.

However, in order to observe $t_h^{\omega+1}$ responses in stratum h, we must sample $t_h^{\omega+1}/\hat{\pi}_h$ units, however, there is no guarantee that there are enough remaining units to sample in the stratum, i.e. that $t_h^{\omega+1}/\hat{\pi}_h \leq N_h - n_h$. A naive approach would simply be to cap the target number of respondents from stratum h at $N_h - n_h$, however this strategy may result in sampling too few units to reach the target total number of responses in the following wave.

$p_h^{\omega+1}$ as function of X

Instead, we can design a strategy to optimize the stratum-level selection probabilities $p_h^{\omega+1}$ to minimize the design effect accounting for the availability of units to sample. Unlike $\hat{\pi}_i$, we are modeling $p_h^{\omega+1}$ at the stratum-level, and thus use stratum covariates $\mathbf{X}_h$. First, we define $p_h^{\omega+1}$ as a function of stratification variables $\mathbf{X}$

$$p_h^{\omega+1} = f(\mathbf{X}_h) = \sigma(\beta^\top X_h) \tag{3.24}$$

where σ is the sigmoid, or inverse logit, function, and β is a vector of regression parameters for each feature of $\mathbf{X}_h$. We next define the expected design effect in terms of $\sigma(\beta^\top X_h)$:

$$\text{deff}(w^{\omega+1}) = 1 + \sum_h (n_h + p_h^{\omega+1}(N_h - n_h)\hat{\pi}_h) \left(\frac{N_h(n + t_h^{\omega+1})}{N(n_h + p_h^{\omega+1}(N_h - n_h)\hat{\pi}_h)} - 1 \right)^2 \tag{3.25}$$

$$= 1 + \sum_h (n_h + \sigma(\beta^\top X_h)(N_h - n_h)\hat{\pi}_h) \left(\frac{N_h(n + t_h^{\omega+1})}{N(n_h + \sigma(\beta^\top X_h)(N_h - n_h)\hat{\pi}_h)} - 1 \right)^2 \tag{3.26}$$

We can then run a gradient-based optimization procedure to estimate the β that will minimize the expected design effect. See Appendix 3.6 for the derivation of gradients of the design effect with respect to β.

Note that $p_h^{\omega+1} = t^{\omega+1}/(N - n_\omega)$ for all h is equivalent to standard stratified random sampling and does not incorporate any information about heterogeneous response rates.

$p_h^{\omega+1}$ for a single outcome of interest

This setting is the most similar to standard active learning with uncertainty sampling or Bayesian Optimization, in which we seek to select a sample in order to minimize posterior uncertainty of a model for Y. In order to implement uncertainty sampling ALSD, on each iteration of ALSD, we first fit a model for Y, with whatever approach we plan to use for analysis of the final sample, for example using MRP. We then calculate predictive uncertainty for each *unobserved* (unlabeled) unit in the population, U_i. If the model is fully Bayesian, U_i could be, for example, the 95% posterior predictive interval, but could also be the predictive standard error or similar. Then, $p_h^{\omega+1}$ is given by:

$$p_h^{\omega+1} = \frac{U_h}{\sum_h U_h} \times t^{\omega+1} \times \frac{1}{N_h - n_h} \times \frac{1}{\hat{\pi}_h} \tag{3.27}$$

where $U_h = \text{median}\{U_i : i \in \mathcal{A}_h\}$. Here the first term represents the relative distribution of posterior uncertainty across all strata, and thus $t^{\omega+1} \times U_h / \sum_h U_h$ gives the target number of respondents per stratum in the next wave. There is no guarantee that $p_h^{\omega+1} \leq 1$, so we cap $p_h^{\omega+1}$ at 1. Similarly, if there are no units remaining to sample in a stratum ($N_h = n_h$), then $p_h^{\omega+1}$ is set to 0.

Unlike in standard active learning, ALSD must account for (often a high degree of) nonresponse as well as the logistical lift of fielding a single survey wave. Therefore, single-outcome ALSD sets $p_h^{\omega+1}$ proportionally to the median stratum-level uncertainty instead of sampling specific units with the highest uncertainties. Stratified sampling also allows for the incorporation of response rate sampling, and, as in batch active learning methods, helps account for similarity between units in the population and ensures that the sample is well-distributed across strata.

Lastly, if small area estimation is the core focus of the survey, we recommend stratifying on the geography of interest which will oversample units from geographies with higher posterior uncertainty.

3.4 Simulation Studies

We study the efficiency of ALSD methods in simulations using survey data collected as part of the World Food Programme's (WFP) mobile Vulnerability Analysis and Mapping (mVAM) project in Zimbabwe (Mock et al., 2016). The data is comprised of 44,599 survey responses collected between 2 September 2020 and 20 May 2023 on mobile and landline telephones. The survey asks a range of demographic questions (e.g. age, gender, head of household education level, type of toilet facilities at home, source of home drinking water, etc) as well as questions intended to gauge the level of food insecurity in the home, which is the primary focus of the survey.

3.4.1 Simulation set up

We compare ALSD sampling methods to two methods commonly used in survey sampling: SRS and quota sampling.

We evaluate methods under each missing at random (MAR) and missing not at random (MNAR) conditions, and across a range of target wave sizes, $t_{j+1} \in (100, 250, 500)$, and true mean response rates, $\mathbb{P}(R_i = 1|S_i = 1) = \pi_i \in (0.05, 0.1, 0.15)$.

On each iteration of the simulation, we generate a true probability of response for each individual in the population, π_i, as a logistic function of features $\mathbf{X}_i$, $\pi_i = \sigma(\beta^\top X_i)$ where $\beta \sim N(0, 3)$. We then adjust the model intercept such that $\bar{\pi}_i$ is equal to the designated true mean response rate for that simulation setting. In order to ensure that there is at least a moderate amount of selection bias present, we generate sets of β_π until $\text{corr}(\pi_i, Y_i) \geq 0.03$.

We set the budget at $B = 1500$ and select an initial simple random sample of 500 units. Then, the remainder of the sample is collected using each of the following methods:

- **ALSD simple**: ALSD in which $t_h^{\omega+1} = \frac{N_h}{N}(n + t^{\omega+1}) - n_h$ (see Section 3.3.5).
- **ALSD modeled**: ALSD in which $p_h^{\omega+1} = \sigma(\beta^\top X_h)$, and β is optimized to minimize the expected deff using a gradient-based procedure (see Section 3.3.5).
- **ALSD MRP**: ALSD in which $p_h^{\omega+1}$ is set according to Equation 3.27. We model Y as a logistic fixed effects model and fit the model using the `glmer` package. U_i is set as the width of the 95% posterior predictive interval.

- **RDD**: WFP mVAM uses random digit dialing (RDD) with quotas as their standard data collection technique. Here we attempt to mimic RDD by specifying a random call order for eligible units, and in that random order, $R_i^{\omega+1} = 1$ if $M_i^{\omega+1} < \pi_i$ where $M_i^{\omega+1} \sim \text{Uniform}(0, 1)$. Like WFP, we set quotas on Administration 1 area and gender and stop observing responses in each quota cell once it is filled.

- **SRS**: Simple random samples where $p_i^{\omega+1} = t^{\omega+1}/\bar{\pi}$ for all units.

Each sampling method is given a maximum of $1 + (B - t_0)/t_{j+1}$ waves to use the allocated budget. Data collection will stop earlier if the budget is exhausted.

In the MAR condition, the true probability of response is a function of gender (2 levels), age (5 levels), home toilet type (3 levels), and Province (10 levels), and these features are all available for modeling $\hat{\pi}_i$, sampling, and weighting. In the MNAR condition, in addition to the 4 features above, the true probability of response is also a function of district (91 levels), but this is not available for modeling, sampling, or weighting.

3.4.2 Results

Table 3.2 shows the median and mean absolute error of weighted estimates of population-wide food insecurity based on samples collected using each method, across simulation conditions. ALSD methods consistently produce samples with the smallest median and mean absolute error, outperforming RDD- and SRS-based samples across all simulation conditions.

In settings with lower response rates (5% and 10%), modeled ALSD tends to produce the most accurate estimates of food insecurity, while simple ALSD performs better in high response rate settings. This relative performance is consistent with a core motivation for modeled ALSD over simple ALSD – the desire to reallocate sample targets to similar strata when particularly hard-to-reach strata have been exhausted. This problem will be more pronounced in low response rate settings, so modeled ALSD should outperform simple ALSD.

Single outcome ALSD using MRP performs the worst across all settings, even under-performing RDD and SRS. This could be due to a number of factors. First, a key strength of MRP is the ability to account for a large number of predictors and interactions, yet in these simulations the outcome model is a simple fixed-effects model with no tuning. In most practical settings, the MRP model on each iteration would likely have more predictors available and more attention placed on variable selection and model tuning. Second, here we compare the error of estimates of food

Nonresponse type	True $\mathbb{P}(R)$	Sampling method	Med Abs Error	Mean Abs Error
MAR	0.05	ALSD MRP	0.0195	0.024
		ALSD modeled	**0.0112**	**0.0141**
		ALSD simple	0.0131	0.0165
		RDD	0.015	0.0191
		SRS	0.0167	0.0187
	0.10	ALSD MRP	0.0252	0.0288
		ALSD modeled	**0.0098**	**0.0118**
		ALSD simple	0.0126	0.0138
		RDD	0.0153	0.0194
		SRS	0.0167	0.018
	0.15	ALSD MRP	0.0317	0.0336
		ALSD modeled	0.0112	0.0136
		ALSD simple	**0.0107**	**0.0115**
		RDD	0.0132	0.0164
		SRS	0.0149	0.018
MNAR	0.05	ALSD MRP	0.0166	0.0217
		ALSD modeled	**0.012**	**0.0149**
		ALSD simple	0.0137	0.0166
		RDD	0.0177	0.0202
		SRS	0.0142	0.02
	0.10	ALSD MRP	0.0216	0.0234
		ALSD modeled	0.0109	**0.0122**
		ALSD simple	**0.009**	0.0127
		RDD	0.0123	0.0158
		SRS	0.0153	0.0173
	0.15	ALSD MRP	0.0286	0.0281
		ALSD modeled	0.0111	0.0125
		ALSD simple	**0.0089**	**0.0111**
		RDD	0.0091	0.0126
		SRS	0.01	0.0129

Table 3.2: This table gives median and mean absolute error for each sampling method across simulation conditions. Simulation conditions vary by nonresponse type (MAR or MNAR), and the true average probability of response in the population (5%, 10%, or 15%).

insecurity derived using rake weights, rather than MRP. This disconnect between method of sampling and method of analysis likely puts MRP at a disadvantage.

Figure 3.1 shows the distribution of design effects calculated using weights from raking the final samples to population targets. ALSD results in lower design effects than either RDD or SRS across all simulation conditions. SRS, which does nothing to account for heterogeneous response rates, consistently has the largest design

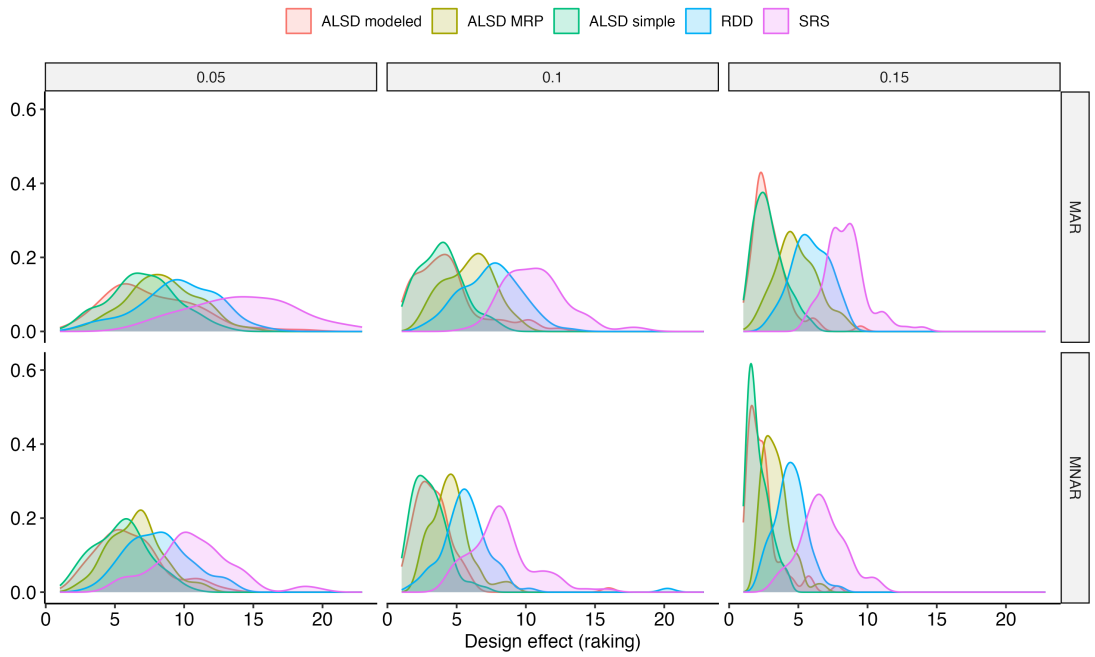


Figure 3.1: The distribution of design effects (deff) calculated for each final sample, weighted using raking to population marginal distributions.

effects. RDD which uses quotas on a limited set of variables has the second largest. ALSD methods, which adaptively design samples based on observed response rates, consistently have the smallest design effects across simulation conditions. Design effects across all methods are lower in higher response rate settings where less weighting is required to correct for heterogeneous nonresponse.

Figure 3.2 compares final sample sizes from each method. Although the overall budget for each run was 1500, some methods were not able to use the entire budget, particularly in difficult response rate settings. A drawback of simple ALSD, as noted in Section 3.3.5 is that it does not account for what population units are actually available to sample when it sets the probability of selection. Modeled ALSD was designed to overcome this drawback, which it appears to in these simulations, and manages to use more of the available budget than other ALSD methods, RDD, and even SRS in some conditions. "Effective n" in Figure 3.2 shows the distribution of *effective* sample size, once the variance inflation from weighting is accounted for. Modeled ALSD maintains its efficiency over other methods, while the larger raw sample sizes of SRS are overcome by the additional variance from weighting.

3.5 Discussion and future work

We propose ALSD as a novel framework for design survey samples. ALSD borrows from Bayesian Optimization and active learning literature to adapt sampling for

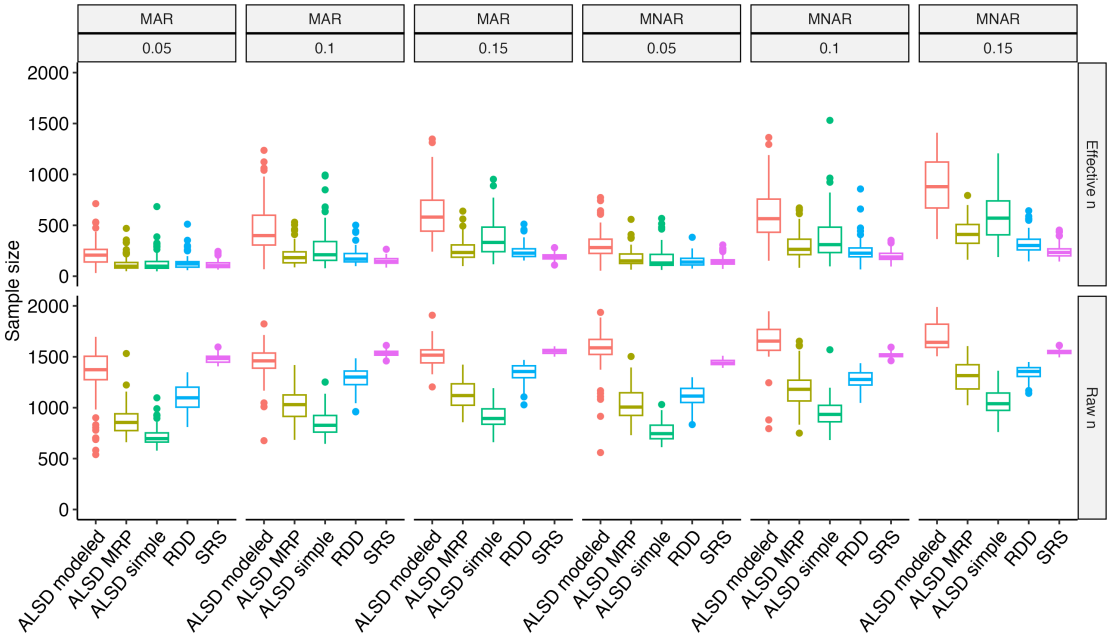


Figure 3.2: The distribution of final Raw and Effective sample sizes. The effective sample size accounts for the additional variance from weighting, $n_{eff} = n/\text{deff}$.

survey research to account for heterogeneous, and even dynamic, response patterns and make the most efficient use of the population available to sample.

Our simulation studies demonstrate the potential efficiency gains from ALSD. ALSD not only consistently produces the lowest absolute error estimates of the key outcome, but the lowest-variance estimates as well.

On December 30, 2023 we concluded a live test of ALSD in Zimbabwe with the WFP, and will evaluate the accuracy and efficiency of ALSD compared to WFP's standard data collection program. We collected 1500 survey responses across 4 waves. The initial sample was a stratified random sample designed to collect 750 total responses. The remaining 3 waves used ALSD to design samples that to collect 250 responses in each wave. Preliminary results show that ALSD was able to adapt to heterogeneous response rates and decrease the design effect of the cumulative sample without utilizing quota sampling. A future paper will discuss these results in greater detail.

Future research should also examine the design choices of ALSD, including the wave size relative to initial sample size, more expressive or nonlinear models of response, and cases in which the covariates are high-dimensional and correspond to a structured domain (as in social network data). Whereas we have opted in this work to model selection probabilities at the stratum level, when they are modeled as a function of the features, one may opt instead to model them at an individual

level. This strategy would be particularly attractive in the case where response probabilities can be accurately modeled at the individual level. ALSD can also be expanded to account for heterogeneous cost and responsiveness by mode of data collection (e.g. CATI, SMS, etc.), and by demographic characteristics of the unit (e.g. rural respondents may be more expensive to observe). Furthermore, we plan to extend ALSD to incorporate spatial and temporal dimensions of survey data.

Appendix

3.6 Appendix: Gradient derivation

Following from Equation 3.26, we seek to derive gradients for the expected design effect with respect to logistic regression parameters β in the model that defines stratum-level selection probability $p_h = \sigma(\beta^\top X_h)$. The optimal selection probabilities $p^{\omega+1}$ must produce an expected value of respondents in the next wave equal to $t^{\omega+1}$, which is a parameter set at initialization.
We introduce some additional notation to simplify further calculations:

- $A = (A_1, ..., A_H)$ is a $(1, H)$ vector with components $A_h = (N_h - n_h^\omega)\pi_h$
- X is the design matrix (H, K); X_h is its h-th row $(1, K)$; X^j is its j-th column $(H, 1)$; X_h^j is element in the h-th row and j-th column
- β is a $(K, 1)$ vector of coefficients in the logistic function of selection probability
- $z_h = \beta^\top X_h$ is a scalar $(1, K)(K, 1) = (1, 1)$
- $\sigma(z) = \frac{1}{1+\exp(-z)}$ is the sigmoid (inverse logit) function
- $\Sigma_h = \sigma(z_h) = \sigma(\beta^\top X_h) = p_h^{\omega+1}$ is a scalar
- $\Sigma = (\Sigma_1, ..., \Sigma_H)^T = (\sigma(\beta^\top X_1), ...\sigma(\beta^\top X_H))^T, \sigma(\beta^\top X) = p^{\omega+1}$ is a $(H, 1)$ vector
- $K_h = A_h\Sigma_h(1 - \Sigma_h)$
- $K = (K_1, ..., K_H)$
- $S_h = n_h^\omega + A_h\Sigma_h$
- $S = n^\omega + A \cdot \Sigma$

In these terms, we get

$$\begin{aligned}
\mathbb{E}[n^{\omega+1}] &= n^\omega + \sum_h (N_h - n_h^\omega) \times \pi_h \times p_h^{\omega+1} \\
&= n^\omega + A \cdot \Sigma, \\
\mathbb{E}[n_h^{\omega+1}] &= n_h^\omega + (N_h - n_h^\omega)\pi_h p_h^{\omega+1} \\
&= n_h^\omega + A_h\Sigma_h
\end{aligned}$$

and 3.26 can be re-written as

$$\text{deff}(w^{\omega+1}) = 1 + \frac{1}{n^\omega + A \cdot \Sigma} \sum_h (n_h^\omega + A_h \Sigma_h) \left(\frac{N_h (n^\omega + A \cdot \Sigma)}{N(n_h^\omega + A_h \Sigma_h)} - 1 \right)^2 \tag{3.28}$$

We have a target budget in mind denoted by $t^{\omega+1}$, which we can go over or under by some tolerance. We think of this as a constraint and turn the constrained optimization problem into an unconstrained optimization problem using Lagrange multipliers:

$$|A^T \Sigma - t^{\omega+1}| \leq \Delta$$

$$\mathcal{L}(\beta) = 1 + \frac{1}{n^\omega + A \cdot \Sigma} \sum_h \underbrace{(n_h^\omega + A_h \Sigma_h) \left(\frac{N_h (n^\omega + A \cdot \Sigma)}{N(n_h^\omega + A_h \Sigma_h)} - 1 \right)^2}_{\mathcal{F}_h(\beta)} + \lambda \underbrace{(A \cdot \Sigma - t^{\omega+1})^2}_{\mathcal{C}(\beta)} \tag{3.29}$$

$$= 1 + \frac{1}{S} \sum_h \underbrace{S_h \left(\frac{N_h S}{N S_h} - 1 \right)^2}_{\mathcal{F}_h(\beta)} + \lambda \underbrace{(A \cdot \Sigma - t^{\omega+1})^2}_{\mathcal{C}(\beta)} \tag{3.30}$$

To derive this expression, let's pre-compute

$$\frac{\partial \Sigma_h}{\partial \beta_j} = \frac{\partial \sigma}{\partial z_h} \frac{\partial z_h}{\partial \beta_j} = \sigma(z_h)(1 - \sigma(z_h)) X_h^j = \Sigma_h (1 - \Sigma_h) X_h^j, \tag{3.31}$$

$$\frac{\partial (A_h \Sigma_h)}{\partial \beta_j} = A_h \sigma(z_h)(1 - \sigma(z_h)) X_h^j = A_h \Sigma_h (1 - \Sigma_h) X_h^j = K_h X_h^j, \tag{3.32}$$

$$\frac{\partial (A \cdot \Sigma)}{\partial \beta_j} = \frac{\partial \sum_i A_i \Sigma_i}{\partial \beta_j} = \sum_i A_i \frac{\partial \Sigma_i}{\partial \beta_j} \tag{3.33}$$

$$= \sum_i A_i \Sigma_i (1 - \Sigma_i) X_i^j = \sum_i K_i X_i^j = K \cdot X^j, \tag{3.34}$$

$$\frac{\partial}{\partial \beta_j} \left(\frac{n^\omega + A \cdot \Sigma}{n_h^\omega + A_h \Sigma_h} \right) = \frac{\frac{\partial (A \cdot \Sigma)}{\partial \beta_j} (n_n^\omega + A_h \Sigma_h) - \frac{\partial (A_h \Sigma_h)}{\partial \beta_j} (n^\omega + A \cdot \Sigma)}{(n_h^\omega + A_h \cdot \Sigma_h)^2} \tag{3.35}$$

$$= \frac{K \cdot X^j S_h - K_h X_h^j S}{S_h^2}, \tag{3.36}$$

$$\frac{\partial}{\partial \beta_j} \left(\frac{1}{n_h^\omega + A_h \Sigma_h} \right) = -\frac{1}{(n^\omega + A \cdot \Sigma)^2} \frac{\partial (A \cdot \Sigma)}{\partial \beta_j} = -\frac{1}{S^2} K \cdot X^j, \tag{3.37}$$

$$\frac{\partial}{\partial \beta_j} \left[\left(\frac{N_h (n^\omega + A \cdot \Sigma)}{N(n_h^\omega + A_h \Sigma_h)} - 1 \right)^2 \right] = 2 \left(\frac{N_h (n^\omega + A \cdot \Sigma)}{N(n_h^\omega + A_h \Sigma_h)} - 1 \right) \frac{N_h}{N} \frac{\partial}{\partial \beta_j} \left(\frac{n^\omega + A \cdot \Sigma}{n_h^\omega + A_h \Sigma_h} \right) \tag{3.38}$$

$$= 2 \left(\frac{N_h S}{N S_h} - 1 \right) \frac{N_h}{N} \frac{K \cdot X^j S_h - K_h X_h^j S}{S_h^2} \tag{3.39}$$

$$\frac{\partial}{\partial\beta_j}\mathcal{F}_h = \frac{\partial}{\partial\beta_j}\left[(n_h^\omega + A_h\Sigma_h)\left(\frac{N_h(n^\omega + A\cdot\Sigma)}{N(n_h^\omega + A_h\Sigma_h)} - 1\right)^2\right]$$
$$= K_h X_h^j \left(\frac{N_h S}{N S_h} - 1\right)^2 + 2\left(\frac{N_h S}{N S_h} - 1\right)\frac{N_h}{N}\frac{(K\cdot X^j)S_h - K_h X_h^j S}{S_h^2} S_h,$$
$$= K_h X_h^j \left(\frac{N_h S}{N S_h} - 1\right)^2 + 2\left(\frac{N_h S}{N S_h} - 1\right)\frac{N_h}{N}\frac{(K\cdot X^j)S_h - K_h X_h^j S}{S_h},$$

Hence,

$$\frac{\partial \text{deff}}{\partial\beta_j} = \frac{\partial}{\partial\beta_j}\left(\frac{1}{S}\sum_h \mathcal{F}_h\right)$$
$$= \underbrace{-\frac{1}{S^2}K\cdot X^j \sum_h \mathcal{F}_h}_{\text{term 1}} + \underbrace{\frac{1}{S}\sum_h\left[K_h X_h^j\left(\frac{N_h S}{N S_h} - 1\right)^2 + 2\left(\frac{N_h S}{N S_h} - 1\right)\frac{N_h}{N}\frac{(K\cdot X^j)S_h - K_h X_h^j S}{S_h}\right]}_{\text{term 2}}$$

$$\frac{\partial\mathcal{C}(\beta)}{\partial\beta_j} = 2(A\cdot\Sigma - t^{\omega+1})\frac{\partial(A\cdot\Sigma)}{\partial\beta_j} = 2(A\cdot\Sigma - t^{\omega+1})K\cdot X^j$$

3.6.1 Matching expected values

Say that we wish to obtain a certain expected number of respondents in the next wave – we set

$$p^{\omega+1}(x) = \frac{\mu_x^{\omega+1}}{(N_x - n_x^\omega)\,\pi(x)}.$$

A desired expected value could be the one that aims to improve the design effect, i.e. we wish

$$\frac{N_x}{N} \approx \frac{n_x^\omega + \mu_x^{\omega+1}}{n^\omega + \mu^{\omega+1}},$$

and hence, one strategy would be to set

$$\mu_x^{\omega+1} = \left(\frac{N_x}{N}\left(n^\omega + \mu^{\omega+1}\right) - n_x^\omega\right)_+,$$

i.e. oversampled stratum would not be sampled at the next wave, and others would try to exactly match the population proportions. But there are two problems with this strategy:

1. There is no guarantee that

$$\left(\frac{N_x}{N}\left(n^\omega + \mu^{\omega+1}\right) - n_x^\omega\right)_+ \leq (N_x - n_x^\omega)\,\pi(x). \tag{3.40}$$

   This requires choosing small enough $\mu^{\omega+1}$, i.e.

$$\mu^{\omega+1} \leq \min_x (N_x - n_x^\omega)\,\pi(x)\frac{N}{N_x} + n_x^\omega\frac{N}{N_x} - n^\omega. \tag{3.41}$$

2. Even if (3.41) is true for all x, it is in fact impossible to ensure the desired expected value for all x simultaneously since

$$\begin{aligned}\mu^{\omega+1} &= \sum_x \left(\frac{N_x}{N}\left(n^\omega + \mu^{\omega+1}\right) - n_x^\omega\right)_+ \\ &\geq \sum_x \left(\frac{N_x}{N}\left(n^\omega + \mu^{\omega+1}\right) - n_x^\omega\right) \\ &= \mu^{\omega+1}\end{aligned}$$

   with equality only if $\mu^{\omega+1}$ is big enough to make $\frac{N_x}{N}\left(n^\omega + \mu^{\omega+1}\right) - n_x^\omega \geq 0$ for all x. Hence, we would require

$$\mu^{\omega+1} \geq \max_x n_x^\omega\frac{N}{N_x} - n^\omega. \tag{3.42}$$

Thus $\mu^{\omega+1}$ must be sandwiched between the two values for the strategy to work... (there could also be a budget for the number of respondents, so that would be an additional upper bound on $\mu^{\omega+1}$)

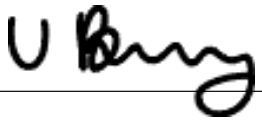
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Adaptive Learning Sampling Design
Publication Status	☐Published ☐ Accepted for Publication ☐Submitted for Publication ☒Unpublished and unsubmitted work written in a manuscript style
Publication Details	**Active Learning Sampling Design (ALSD)** **Valerie C. Bradley, Elizaveta Semenova, Adam Howes, Mengyan Zhang, Theo Rashic, Jeff Imai-Eaton, Dino Sejdinovic, and Seth Flaxman**

Student Confirmation

Student Name:	Valerie C Bradley		
Contribution to the Paper	Jointly developed the idea with S.F.; jointly developed the theory with S.F., D.S., M.Z. and E.S.; performed all simulations and statistical analysis, wrote manuscript		
Signature		Date	Jan 5, 2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: **Prof Dino Sejdinovic**		
Signature 	Date	7 January 2024

This completed form should be included in the thesis, at the end of the relevant chapter.

4
Weighting Leverage

Abstract

Inverse probability weighting (IPW) is commonly used to adjust for nonresponse in surveys. IPW methods adjust observations using a set of auxiliary variables $\mathbf{X}$ such that the weighted distribution of $\mathbf{X}$ in the observations matches known population distributions $p_{\mathbf{X}}$. IPW estimators will be unbiased if two key assumptions hold: 1) conditional ignorability of outcome Y and response mechanism R, i.e. $Y \perp\!\!\!\perp R|\mathbf{X}$, and 2) that $p_{\mathbf{X}}$ is known. The conditional ignorability assumption is widely studied, but the second assumption much less so. In practice, $p_{\mathbf{X}}$ is often estimated from other surveys, and the uncertainty and potential bias in its estimation ignored. For example, we know that weighting on education level is critical for ensuring accuracy of US pre-election polls, however there remains disagreement about the true level of education in the 2016 and 2020 electorates, and considerable uncertainty in the level of education of the future 2024 electorate.

Here we introduce *leverage* of auxiliary variables, which measures the sensitivity of the weighted estimator of a population mean $\hat{Y}_w$ to error in population weighting targets $p_{\mathbf{X}}$. Leverage can also be used to assess the relative sensitivity of a population mean to elements in a set of auxiliary variables. We derive *leverage* by decomposing the bias of $\hat{Y}_w$ that results from an incorrect population target, provide estimation methods, and assess these methods in simulation studies. Finally, we demonstrate *leverage*'s usefulness in practice for incorporating uncertainty of population targets into weighted survey estimates with an application to the Axios-Ipsos Coronavirus Tracker.

4.1 Introduction

Selection bias occurs when the mechanism that governs which units from a population are observed is correlated with an outcome of interest (Little and Rubin, 2019). In survey research, selection bias often takes the form of nonresponse bias, when units selected for a survey fail to respond to a survey invitation. Selection bias can also occur in surveys due to lack of coverage in the sampling frame, or due to choices made during analysis to exclude particular observations. If not properly addressed, selection bias can lead to large estimation error or misleading conclusions in survey-based estimates. Pre-election polls in 2016 and 2020 US presidential contests, the 2016 Brexit referendum in the UK, and surveys measuring COVID vaccine uptake in the US in 2021 are examples of selection bias in the wild with high practical impact (Jackson, 2016; Kennedy et al., 2018; Clinton et al., 2020; Bradley et al., 2021).

One of the most common methods of addressing selection bias in survey research is inverse probability weighting (IPW) (Horvitz and Thompson, 1952b; Deville and Särndal, 1992). IPW methods generate weights for observed units using a set of auxiliary variables $\mathbf{X}$ such that the weighted distribution of $\mathbf{X}$ in the sample matches the population distribution of $\mathbf{X}$, $p_{\mathbf{X}}$. In order for IPW to successfully eliminate the impact of selection bias on sample-based estimates of an outcome Y, (at least) two key assumptions must be met. First, the set $\mathbf{X}$ must be chosen such that an outcome Y is *conditionally independent* of the sample inclusion indicator R, i.e. $Y \perp\!\!\!\perp R|\mathbf{X}$, where $R_i = 1$ when unit i is observed in the sample, and 0 otherwise. Second, we assume that $p_{\mathbf{X}}$ is known, or at least estimated accurately. However this second assumption is often not articulated explicitly. A core tension that survey researchers face in selecting $\mathbf{X}$ is that $\mathbf{X}$ should contain the full set of features needed to induce conditional independence, however if $\mathbf{X}$ is too high dimensional, estimation of $p_{\mathbf{X}}$ becomes more difficult and certain IPW methods, like post-stratification, become unstable.

There is a large existing literature on methods for selecting $\mathbf{X}$ in order to satisfy the first assumption. Geuzinge et al. (2000); Bethlehem and Schouten (2004); Caughey and Hartman (2017) propose a series of outcome-dependent metrics for ranking elements (or even subsets) of $\mathbf{X}$ by their importance in adjustment while Särndal (2008) introduces an outcome-agnostic metric. Hartman et al. (2021) approach the problem as one of dimensionality reduction rather than selection of the optimal subset, and Hartman and Huang (2024) develop a method for testing the sensitivity of population estimates to the correct selection of the auxiliary set. Furthermore, we know that the importance of certain auxiliary variables can change

over time, as discussed by (Kennedy et al., 2018) in the context of the changing importance of education weighting in political surveys from 2012 to 2016.

The literature addressing the second assumption, that of certainty in the population weighting targets, is more sparse. Population frames for survey weighting are often constructed or modeled from multiple data sources (Leemann and Wasserfallen, 2017; Lauderdale et al., 2020; Kuriwaki et al., 2023). Even marginal weighting targets that are generated from large "gold standard" surveys (e.g. in American political applications, the Cooperative Congressional Election Study or the American National Election Study) have their own associated uncertainty. While population frame uncertainty is a well-documented component of the Total Survey Error framework (Groves and Lyberg, 2010), and widely known to be a risk to survey accuracy, methods for assessing and quantifying that risk are not readily available. Prior attempts to capture population frame uncertainty have involved simulation (Lauderdale et al., 2020), or the introduction of "benchmark uncertainty intervals" (Bradley et al., 2021), however large-scale simulation may be out-of-reach for the average survey researcher, and the "benchmark imprecision intervals" as introduced thus far lack strong theoretical grounding.

When either assumption is violated, selection bias will persist in survey-based estimates of population quantities. Meng (2018) develops a mathematical framework for quantifying selection bias, which we will use throughout this work. This framework decomposes error in an estimator of a population mean into 3 components: data quantity, population heterogeneity, and data quality measured as the *data defect correlation* (*ddc*). See Appendix 4.7 for a more in-depth explanation.

In this work, we introduce the metric *leverage* as another tool to help address these challenges. *Leverage* analytically measures the sensitivity of the weighted estimate of Y to the adjustment target (e.g. estimated population mean) of a particular auxiliary variable. When the absolute value of leverage of an auxiliary variable for an outcome is high, small changes in the sample distribution of that variable drive large changes in the weighted estimate of the outcome, and are therefore critical to adjust for. Leverage can be estimated without population data for $\mathbf{X}$, so can be used to prioritize acquisition of population benchmark data, and to continuously monitor the importance of auxiliary variables for adjustment as the response landscape changes. We also show how leverage can be used to construct confidence intervals for weighted estimators that account for uncertainty in a weighting benchmark.

The paper proceeds as follows. Section 4.2 outlines the theoretical basis for the leverage metric and methods for estimation in practice. Section 4.3 presents

simulation studies exploring methods for estimating leverage and their robustness to key assumptions. Finally, Sections 4.4 demonstrates uses of leverage in practice on the Axios-Ipsos Coronavirus Survey.

4.2 Leverage of auxiliary variables

We propose a new metric *leverage* for capturing the sensitivity of the weighted estimator for a quantity Y to the population weighting target of a binary auxiliary variable U, conditional on a set of additional auxiliary variables $\mathbf{X}$, which we will call control features. Leverage quantifies the impact of error in the estimate of the marginal distribution of U, on the bias of the weighted estimator for the population mean of Y, assuming that $Y \perp\!\!\!\perp R|\{\mathbf{X}, U\}$ and that $p_{\mathbf{X}}$ is known.

In this section, we first introduce notation and the problem setting, then decompose bias in an estimator of a population mean to capture the impact of error in population frame targets. We then define *leverage* and discuss three estimation methods: regression, perturbation, and conditional expectation.

4.2.1 Notation and problem setting

First, we establish some notation for the nonresponse adjustment setting. Let us assume a finite population of units $i = \{1, \ldots, N\}$, of which a subset of size n are observed in a survey ($R_i = 1$ when unit i is observed, 0 otherwise). We would like to estimate population mean $\bar{Y}_N$, but only observe Y_i for units where $R_i = 1$. We assume *conditional ignorability* (Little and Rubin, 2019) of Y and R given a set of auxiliary variables $\{\mathbf{X}, U\}$, such that $Y \perp\!\!\!\perp R|\mathbf{X}, U$.

We observe a *survey* in which we collect the outcome and auxiliary information from respondents, $(Y_i, U_i, \mathbf{X}_i) \sim p_{Y,U,X|R=1}$. We also observe auxiliary information about the population $(U_i, \mathbf{X}_i) \sim p_{U,X}$, which we will call a *census*, though it is often estimated from other surveys or datasets rather than from a true census. To estimate $\bar{Y}_N$, we use an IPW estimator to correct for nonresponse bias:

$$\bar{Y}_w = \frac{\sum_R Y_i w_i}{n} \tag{4.1}$$

where the sum over R indicates a sum over units in the sample, and w_i is the estimated IPW weight for sample unit i. Weights for units where $R_i = 1$ are given by:

$$w_i \propto \frac{\mathbb{P}(R_i = 1)}{\mathbb{P}(R_i = 1|\mathbf{x}_i, u_i)} = \frac{\mathbb{P}(\mathbf{x}_i, u_i)}{\mathbb{P}(\mathbf{x}_i, u_i|R_i = 1)}. \tag{4.2}$$

we will assume here that weights have been normalized such that $\sum_R w_i = n$.

$\bar{Y}_w$ will be an unbiased estimator of $\bar{Y}_N$ when the assumption of conditional ignorability of response holds. This assumption leads exactly to post-stratification weights, however post-stratification becomes difficult when the number of strata defined by the joint distribution of $\{\mathbf{X}, U\}$ becomes large. In this case, calibration weighting may be more appropriate (Deville and Särndal, 1992). The assumption of *linear* ignorability (Hartman et al., 2021) generalizes the conditional ignorability assumption to justify a larger set of weighting algorithms that rely on more complex functions of auxiliary variables than those used in post-stratification, e.g. $\phi(\mathbf{X}, U)$. However, in practice, it is impossible to confirm that either assumption holds, so we must rely on contextual knowledge to assess vulnerability to violations of these assumptions. As we will see, leverage will be a useful tool for such assessment.

In many practical cases, survey researchers do not have access to the exact $p_{\mathbf{X},U}$ to use in IPW weighting. For example, in American pre-election polling, the true population of voters in an upcoming election is unknown, and instead, $p_{\mathbf{X},U}$ is *estimated* from turnout in prior similar elections, high-quality government surveys, and contextual knowledge. This *population frame uncertainty* is a large potential source of bias in pre-election polls.

Leverage seeks to quantify the impact on the bias of $\bar{Y}_w$ from using $p_{\mathbf{X},\tilde{U}}$ in IPW weighting, instead of $p_{\mathbf{X},U}$, where $\tilde{U}$ is a noisy or biased estimate of U. For simplicity, we assume that $U \sim \text{Bern}(p)$ and $\tilde{U} \sim \text{Bern}(\tilde{p})$. Finally, let $q(\mathbf{x}) = \mathbb{E}[U|\mathbf{X} = \mathbf{x}]$, $\tilde{q}(\mathbf{x}) = \mathbb{E}[\tilde{U}|\mathbf{X} = \mathbf{x}]$, and $f(\mathbf{x}, u) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, U = u]$.

4.2.2 Bias in $\bar{Y}_n$ from $\tilde{U}$

To introduce *leverage*, we first focus on decomposing the bias in the estimate of a population mean that results from using $\tilde{U}$ rather than U in adjustment. We begin in the post-stratification weighting setting, and decompose the weights given in Equation 4.2 into 2 components:

$$w_i = \frac{\mathbb{P}(\mathbf{x}_i, u_i)}{\mathbb{P}(\mathbf{x}_i, u_i|R_i = 1)} = \frac{\mathbb{P}(\mathbf{x}_i)}{\mathbb{P}(\mathbf{x}_i|R_i = 1)} \times \frac{\mathbb{P}(u_i|\mathbf{x}_i)}{\mathbb{P}(u_i|\mathbf{x}_i, R_i = 1)} = w_{1,i} \times w_{2,i}$$

The weighted sample mean $\bar{Y}_w = \sum_R Y_i w_{1i} w_{2i}/n$ is unbiased because $Y \perp\!\!\!\perp R|\mathbf{X}, U$.

However, say we construct weights with $p_{\mathbf{X},\tilde{U}}$ instead of $p_{\mathbf{X},U}$. In this case, our estimator is actually

$$\bar{Y}_{\tilde{w}} = \frac{1}{n} \sum_R Y_i w_{1i} \tilde{w}_{2i} \tag{4.3}$$

where $\tilde{w}_{2,i} = \frac{\mathbb{P}(\tilde{u}_i|\mathbf{x}_i)}{\mathbb{P}(u_i|\mathbf{x}_i, R_i=1)}$ is the incorrect weight term. Note, however, that the denominator of $\tilde{w}$ remains correct and is given in terms of u_i (rather than $\tilde{u}_i$),

because we assume that we have measured the correct quantity U for respondents, but not in the wider population. For example, this would occur in practice if the educational attainment of respondents was collected directly in the survey (U), but population educational attainment targets are based on modeled estimates that are incorrectly calibrated ($\tilde{U}$).

The expected value of $\bar{Y}_{\tilde{w}}$ is given by:

$$\mathbb{E}[\bar{Y}_{\tilde{w}}] = \sum_{\mathbf{x},u} \mathbb{E}_Y \left[Y \frac{\mathbb{P}(\mathbf{x})}{\mathbb{P}(\mathbf{x}|R=1)} \frac{\mathbb{P}(\tilde{u}|\mathbf{x})}{\mathbb{P}(u|\mathbf{x},R=1)} | \mathbf{x}, u, R=1 \right] \mathbb{P}(\mathbf{x}, u|R=1) \tag{4.4}$$

$$= \sum_{\mathbf{x},u} \mathbb{E}_Y \left[Y|\mathbf{x}, u, R=1 \right] \frac{\mathbb{P}(\mathbf{x}, \tilde{u})}{\mathbb{P}(\mathbf{x}, u|, R=1)} \mathbb{P}(\mathbf{x}, u|R=1) \tag{4.5}$$

$$= \sum_{\mathbf{x},u} \mathbb{E}_Y \left[Y|\mathbf{x}, u, R=1 \right] \mathbb{P}(\mathbf{x}, \tilde{u}). \tag{4.6}$$

Assuming conditional ignorability, $Y \perp\!\!\!\perp R|\{\mathbf{X}, U\}$, we can further simplify:

$$\mathbb{E}[\bar{Y}_{\tilde{w}}] = \sum_{\mathbf{x},u} \mathbb{E}_Y \left[Y|\mathbf{x}, u \right] \mathbb{P}(\mathbf{x}, \tilde{u}) \tag{4.7}$$

$$= \sum_{\mathbf{x},u} f(\mathbf{x}, u) \mathbb{P}(\mathbf{x}, \tilde{u}) \tag{4.8}$$

When U is binary,

$$\mathbb{E}[\bar{Y}_{\tilde{w}}] = \sum_{\mathbf{x}} f(\mathbf{x}, 1)\mathbb{P}(\mathbf{x}, \tilde{U}=1) + f(\mathbf{x}, 0)\mathbb{P}(\mathbf{x}, \tilde{U}=0) \tag{4.9}$$

$$= \sum_{\mathbf{x}} \Big(f(\mathbf{x}, 1)\mathbb{P}(\tilde{U}=1|\mathbf{x}) + f(\mathbf{x}, 0)(1 - \mathbb{P}(\tilde{U}=1|\mathbf{x})) \Big) \mathbb{P}(\mathbf{x}) \tag{4.10}$$

$$= \sum_{\mathbf{x}} \left(f(\mathbf{x}, 1)\tilde{q}(\mathbf{x}) + f(\mathbf{x}, 0)(1 - \tilde{q}(\mathbf{x})) \right) \mathbb{P}(\mathbf{x}) \tag{4.11}$$

Hence, the bias in $\bar{Y}_{\tilde{w}}$ is:

$$\text{Bias}(\bar{Y}_{\tilde{w}}) = \mathbb{E}\left[\bar{Y}_{\tilde{w}}\right] - \bar{Y}_N \tag{4.12}$$

$$= \sum_{\mathbf{x}} \left(f(\mathbf{x}, 1)\tilde{q}(\mathbf{x}) + f(\mathbf{x}, 0)(1 - \tilde{q}(\mathbf{x})) - f(\mathbf{x}, 1)q(\mathbf{x}) - f(\mathbf{x}, 0)(1 - q(\mathbf{x})) \right) \mathbb{P}(\mathbf{x}) \tag{4.13}$$

$$= \sum_{\mathbf{x}} \left(f(\mathbf{x}, 1) - f(\mathbf{x}, 0) \right) \left(\tilde{q}(\mathbf{x}) - q(\mathbf{x}) \right) \mathbb{P}(\mathbf{x}) \tag{4.14}$$

$$= \mathbb{E}_{\mathbf{X}} \left[\left(f(\mathbf{x}, 1) - f(\mathbf{x}, 0) \right) \left(\tilde{q}(\mathbf{x}) - q(\mathbf{x}) \right) \right] \tag{4.15}$$

The bias in the weighted estimator of $\bar{Y}$ that uses $\tilde{U}$ rather than U is the expected value, relative to the population strata defined by control features $\mathbf{X}$, of the product of $f(\mathbf{x}, 1) - f(\mathbf{x}, 0)$, the difference between the conditional mean of Y given $\mathbf{x}$ where $U = 1$ and the same where $U = 0$, and $\tilde{q}(\mathbf{x}) - q(\mathbf{x})$, which captures the degree of error in our estimate of the population rate $\mathbb{E}(U|\mathbf{x})$ from using a noisy estimate, $\tilde{U}$.

4.2.3 Estimating Leverage

Estimating bias directly using the decomposition above requires that the components capturing the magnitude of distributional error, $\tilde{q}(\mathbf{x}) - q(\mathbf{x})$, and sensitivity of the outcome to that distributional error, $f(\mathbf{x}, 1) - f(\mathbf{x}, 0)$, be calculated for each value of $\mathbf{x}$. In this section we introduce three methods for estimating *leverage*, a single metric summarizing the sensitivity: regression, conditional expectation, and perturbation.

Regression

The simplest method for summarizing sensitivity of $\bar{Y}_n$ to $\tilde{U}$ uses simple linear regression. If we assume that Y is continuous and $f(\mathbf{x}, u)$ is a well-specified (and unstandardized) linear regression model

$$f(\mathbf{x}, u) = \beta_0 + \beta_u u + \beta_x^T \mathbf{x} \tag{4.16}$$

$f(\mathbf{x}, 1) - f(\mathbf{x}, 0)$ is simply β_u and is independent of $\mathbf{X}$. Therefore, the bias decomposition simplifies to

$$\begin{aligned}
\text{Bias}(\bar{Y}_{\tilde{w}}) =& \mathbb{E}_{\mathbf{x}} \left[\beta_u \left(\tilde{q}(\mathbf{x}) - q(\mathbf{x}) \right) \right] && (4.17)\\
=& \beta_u \times \mathbb{E}_{\mathbf{x}} \left[\tilde{q}(\mathbf{x}) - q(\mathbf{x}) \right] && (4.18)\\
=& \beta_u \times \left(\mathbb{E}(\tilde{U}) - \mathbb{E}(U) \right) && (4.19)\\
=& \beta_u^* \times \sigma_Y \times \frac{\mathbb{E}(\tilde{U}) - \mathbb{E}(U)}{\sigma_U} && (4.20)\\
=& \lambda_U \times \sigma_Y \times \delta_U && (4.21)
\end{aligned}$$

where λ_U is what we will refer to as *leverage* in this case, β_u^* is the standardized regression coefficient from $f(\mathbf{x}, u)$, and δ_U is defined as $\frac{\mathbb{E}(\tilde{U}) - \mathbb{E}(U)}{\sigma_U}$.

Therefore, in the linear case, the bias in a weighted sample mean $\bar{Y}_w$ from using a noisy or biased weighting target for auxiliary variable U can be decomposed into the following readily interpretable quantities:

- **Leverage of** U given $\mathbf{X}$, $\lambda_U \coloneqq \beta_u^*$. The leverage of U measures the sensitivity in outcome Y to auxiliary variable U, given a set of control features $\mathbf{X}$. Estimator bias will increase with the magnitude of the partial correlation of Y and U, controlling for $\mathbf{X}$.

- **Distribution error**, $\delta_U \coloneqq \frac{\mathbb{E}(\tilde{U}) - \mathbb{E}(U)}{\sigma_U}$. The distribution error measures the standardized magnitude of error in the marginal distribution of $\tilde{U}$. Estimator bias will increase with the magnitude of error in the weighting target.

- **Population heterogeneity**, σ_Y. We use the name for this quantity from Meng (2018). The potential for bias will increase with the population variance of the outcome.

In practice, we rarely observe the bias in an estimator or the magnitude of error in a weighting target. Furthermore, it is important to note that σ_Y is the *population* standard deviation of Y, which may not be estimable from the observed sample.

Estimating leverage using simple linear regression can be done in closed-form, but requires strong assumptions about the functional form of f. If instead $f(\mathbf{x}, u)$ is of the form $f(\mathbf{x}, u) = \beta_0 + \beta_u u + \beta_x \mathbf{x} + \beta_{ux} u\mathbf{x}$, then $f(\mathbf{x}, 1) - f(\mathbf{x}, 0) = \beta_u + \beta_{ux}\mathbf{x}$, which is clearly not independent of X nor estimable as a single regression coefficient. The assumption that $f(\mathbf{x}, u)$ lacks interaction terms between U and elements of $\mathbf{X}$ is not uncommon in survey statistics. In fact, it is the same assumption made by the IPW method raking (Deville and Särndal, 1992). The next section discusses an alternative approach to estimating leverage that, while not closed-form, requires less-stringent assumptions.

Conditional Expectation

We can establish a more general notion of leverage by instead assuming that the difference in conditional means, $f(\mathbf{x}, 1) - f(\mathbf{x}, 0)$, is independent of the distributional error $\tilde{q}(\mathbf{x}) - q(\mathbf{x})$. In this case, from Equation 4.15:

$$
\begin{aligned}
Bias(\bar{Y}_{\tilde{w}}) =& \mathbb{E}_{\mathbf{X}}\left[f(\mathbf{x}, 1) - f(\mathbf{x}, 0)\right] \mathbb{E}_{\mathbf{X}}\left[\tilde{q}(\mathbf{x}) - q(\mathbf{x}))\right] && (4.22)\\
=& \mathbb{E}_{\mathbf{X}}\left[f(\mathbf{x}, 1) - f(\mathbf{x}, 0)\right] \times \Big(\mathbb{E}(\tilde{U}) - \mathbb{E}(U)\Big) && (4.23)\\
=& \mathbb{E}_{\mathbf{X}}\left[f(\mathbf{x}, 1) - f(\mathbf{x}, 0)\right] \times \Big(\mathbb{E}(\tilde{U}) - \mathbb{E}(U)\Big) \times \frac{\sigma_U}{\sigma_U} && (4.24)\\
=& \mathbb{E}_{\mathbf{X}}\left[f(\mathbf{x}, 1) - f(\mathbf{x}, 0)\right] \times \delta_U \times \sigma_U \times \frac{\sigma_Y}{\sigma_Y} && (4.25)\\
=& \left(\frac{\sigma_U}{\sigma_Y} \mathbb{E}_{\mathbf{X}}\left[f(\mathbf{x}, 1) - f(\mathbf{x}, 0)\right]\right) \times \sigma_Y \times \delta_U. && (4.26)
\end{aligned}
$$

Thus, in general, leverage is defined as:

$$
\lambda_U := \frac{\sigma_U}{\sigma_Y} \mathbb{E}_{\mathbf{X}}\left[f(\mathbf{x}, 1) - f(\mathbf{x}, 0)\right]. \tag{4.27}
$$

In order to define leverage on the same scale as the standardized regression coefficient, we scale the expectation of difference in conditional means by the standard deviations of Y and U. While slightly less straightforward, this will allow us to compare leverage more easily across U. The distribution error δ_U and population heterogeneity σ_Y are defined the same as in the regression estimator.

To estimate this directly, we need to estimate $f(\mathbf{x}, u)$, or the expectation of Y conditional on $\mathbf{X}$ and U. The simplest estimator for $f(\mathbf{x}, u)$ is the stratum mean, but this estimator may be unstable if the strata defined by the joint distribution of $\mathbf{X}$ and U are sparse in the sample. A better approach would be to estimate the conditional means with an appropriately specified model, calculate $f(\mathbf{x}, 1) - f(\mathbf{x}, 0)$ for each value of $\mathbf{x}$, and integrate over $\mathbf{x}$. This method is quite similar to performing MRP (Park et al., 2004, 2006) where the outcome is $f(\mathbf{x}, 1) - f(\mathbf{x}, 0)$ and population frame defined by $\mathbf{x}$ alone. While this method can accommodate a wider array of functional forms of f, it still requires that f be correctly specified.

Perturbation

Alternatively, we can estimate leverage without making any assumptions about the form of f with *perturbation*. With this approach, we treat available estimates of p_U as if they are correct, then perturb p_U in order to estimate the effect on Y_n. We first use any IPW method to estimate w using $P(\mathbf{X})$ and $\mathbb{E}(U)$, then use the same IPW method to estimate a second set of weights $\tilde{w}$ using $P(\mathbf{X})$ and $\mathbb{E}(\tilde{U})$, where $\mathbb{E}(\tilde{U}) = \mathbb{E}(U) + \delta_U \sigma_U$ for some chosen value of δ_U. Then, we estimate leverage simply as:

$$\hat{\lambda}_U = \frac{\bar{Y}_{\tilde{w}} - \bar{Y}_w}{\sigma_Y \times \delta_U} \tag{4.28}$$

This method most closely approximates the true practical impact of error in a population target used in weighting as it will reflect the actual change in the weighted mean of a specific sample given the particular set of choices made by the analyst (e.g. weighting method).

Perturbation assumes that Y is *locally* linear in U, with a range determined by δ_U, therefore estimates of λ will depend on the choice of δ_U. Recall from 4.2.3 that δ_U quantifies the error in the marginal distribution of estimated of U. With perturbation, instead of measuring existing error in the marginal distribution of U, we use δ_U to introduce error, then measure the resulting change in $\bar{Y}_n$.

There are a few natural choices for δ_U. First, if there is a specific alternative hypothesis in mind for $\mathbb{E}(\tilde{U})$, then δ_U can be set to $(\mathbb{E}(\tilde{U}) - \mathbb{E}(U))/\sigma_U$. For example, say that our alternative hypothesis for $\mathbb{E}(\tilde{U})$ is the weighted sample mean of U using weights estimated using only the control feature set $\mathbf{X}$. Thus leverage captures the impact of weighting only with $\mathbf{X}$ and not U. We could also set δ_U to be some constant. δ_U measures the number of standard deviations that $\mathbb{E}(\tilde{U})$ lies from $\mathbb{E}(U)$, so, for example, we could set $\delta_U = 1$ to capture the impact of a σ_U-sized error in $\mathbb{E}(U)$.

Section 4.3 explores robustness of estimation methods to key assumptions.

4.2.4 Using leverage in practice

There are a number of uses for *leverage* in practice. First, given a set of survey responses, a researcher might use leverage to rank sensitivity to population target error across potential weighting variables. Given that accurate population benchmark data can be difficult and costly to acquire or estimate well, this can help prioritize population data acquisition. Furthermore, it is possible to estimate leverage for features measured in a survey for which one lacks population data entirely. For example, a political poll may ask a question intended to gauge the level of political engagement of respondents. While the true population distribution of political engagement is incredibly difficult, if not impossible, to estimate well, one could use leverage to identify features that are highly influential on an outcome of interest, and use leverage to help diagnose any latent selection bias that remains after adjustment using standard auxiliary variables.

Leverage is also useful for incorporating population frame uncertainty into estimates of Y. For example, if $\mathbb{E}(U)$ is itself estimated from a survey, then we can use the sampling distribution of U to construct a confidence interval for $\mathbb{E}(U)$. Then, we use leverage to evaluate the impact of uncertainty in $\mathbb{E}(U)$ on $\bar{Y}_w$. The population frame uncertainty interval for $\bar{Y}_w$ is

$$\bar{Y}_w \pm |\lambda_U| \times \frac{\sigma_Y}{\sqrt{m_{\text{eff}}}} \times Z_{1-\alpha_U/2} \tag{4.29}$$

where m_{eff} is the effective sample size of the survey used to estimate $\mathbb{E}(U)$ and $Z_{1-\alpha_U/2}$ is the α_U-level z-score. See Appendix 4.8 for the derivation of these *population frame uncertainty intervals*.

Consider, for example, US pre-election polling. It is well known that weighting on education is critical for accuracy, but the level of education in a future electorate (i.e. the population weighting target), is always unknown. These leverage-based population frame uncertainty intervals are useful tools for quantifying the impact of the uncertainty in the education level of the electorate when presenting weighted estimates of candidate support.

4.3 Simulation studies

In this section we use simulation studies to explore leverage estimation procedures, regression and perturbation, and evaluate their robustness to core assumptions.

4.3.1 Simulation Set-Up

In order to evaluate leverage in the context of a complex correlation structure that adequately replicates dependence between adjustment covariates that we would expect to observe in real survey data, we base our simulations on data pooled from waves 35 through 57 of the Axios-Ipsos Coronavirus Tracker (Jackson et al., 2021), fielded between January and November of 2021 (23,971 survey responses in total). We use these waves because they are the waves that ask about COVID vaccine uptake. For each unit in this population, we observe a set of characteristics including educational attainment, race, age, gender, and urbanicity, and whether each individual has received at least one dose of a COVID vaccine. We use the Axios-Ipsos data as our population, but will generate synthetic outcomes and response mechanisms, described below.

We evaluate leverage estimation methods in four settings that vary in type of missingness (MAR, MNAR) and functional form of $f(\mathbf{x}, u)$. In all settings, we seek to estimate the leverage of the binary variable U indicating whether an individual has a Bachelor's degree on an outcome Y, conditional upon a set of control features $\mathbf{X} = \{X_{\text{under 40}}, X_{\text{white}}\}$, where $X_{\text{under 40}} = 1$ if a respondent is under 40, 0 otherwise, and $X_{\text{white}} = 1$ if a respondent is white, 0 otherwise. In missing at random (MAR) settings where it is possible to recover from nonresponse, the outcome Y and response mechanism R are conditionally independent given $\mathbf{X}$ and U, $Y \perp\!\!\!\perp R|\{\mathbf{X}, U\}$. In practice, survey researchers generally assume that missingness is MAR, however in order to test the robustness of leverage to this assumption, we also examine missing not at random MNAR settings. In MNAR settings, Y and R also depend on an indicator for whether a respondent identifies as a Republican, Z_{repub}, but Z_{repub} is not available for adjustment.

In each setting, we generate Y_i and R_i for each unit in the population as follows:

$$\mu_i = \mathbb{E}[Y_i|\mathbf{X}_i, U_i, Z_i] = f(\mathbf{x}_i, u_i, z_i), \tag{4.30}$$

$$\gamma_Y^2 \sim U(0, 1.5), \tag{4.31}$$

$$Y_i \sim N(\mu_i, \gamma_Y^2), \tag{4.32}$$

$$\pi_i = p(R_i = 1) = \sigma(g(\mathbf{x}_i, u_i, z_i)), \tag{4.33}$$

$$R_i \sim \text{Bernoulli}(\pi_i), \tag{4.34}$$

where σ is the inverse logistic function, and γ controls the strength of the relationship between Y and the auxiliary variables. We set intercepts in each f and g such that the population mean of Y is 0 and the average response rate is 5%. In each setting, coefficients are selected to ensure a severe selection bias in the resulting

samples as measured by *ddc*, or the correlation between Y and R, $\rho_{Y,R}$. We define severe selection bias in terms of the relative reduction in bias-adjusted effective sample size n^*_{eff} (Meng, 2018):

$$1 - \frac{n^*_{\text{eff}}}{n} = 1 - \frac{n}{N-n} \times \frac{1}{\mathbb{E}[\rho^2_{Y,R}]} \times \frac{1}{n} \tag{4.35}$$

Given that our population size for these simulations is $N = 23971$, and our average sample size is 5% of that, or 1200, a 99.9% reduction in effective sample size occurs when $\rho_{Y,R} \approx 0.04$.

The specific functions used to generate μ_i and π_i in each setting are given below:

- **Setting 1**: Missing at random (MAR), simple linear $f(\mathbf{x}, u)$

$$\begin{aligned} f(\mathbf{x}_i, u_i, z_i) &= 0.1 + 0.2X_{i,\text{white}} - 0.1X_{i,\text{under 40}} - 0.5U_i \\ g(\mathbf{x}_i, u_i, z_i) &= -3.2 + 0.5X_{i,\text{white}} - 1.5X_{i,\text{under 40}} + U_i \end{aligned}$$

- **Setting 2**: Missing not at random (MNAR), simple linear $f(\mathbf{x}, u)$

$$\begin{aligned} f(\mathbf{x}_i, u_i, z_i) &= 0.2X_{i,\text{white}} - 0.1X_{i,\text{under 40}} - 0.5U_i \\ &\quad + 0.25Z_i \\ g(\mathbf{x}_i, u_i, z_i) &= -3.1 + 0.5X_{i,\text{white}} - 1.5X_{i,\text{under 40}} + U_i - 0.5Z_i \end{aligned}$$

- **Setting 3**: Missing at random (MAR), $f(\mathbf{x}, u)$ with interactions

$$\begin{aligned} f(\mathbf{x}_i, u_i, z_i) &= 0.1 + 0.2X_{i,\text{white}} - 0.1X_{i,\text{under 40}} - 0.5U_i \\ &\quad - 0.5(1-U_i)X_{i,\text{under 40}} + 1.5U_iX_{i,\text{under 40}} + 0.4U_iX_{i,\text{under 40}}X_{i,\text{white}} \\ g(\mathbf{x}_i, u_i, z_i) &= -3.2 + 0.5X_{i,\text{white}} - 1.5X_{i,\text{under 40}} + U_i \end{aligned}$$

- **Setting 4**: Missing not at random (MNAR), $f(\mathbf{x}, u)$ with interactions

$$\begin{aligned} f(\mathbf{x}_i, u_i, z_i) &= 0.3 + 0.2X_{i,\text{white}} - 0.1X_{i,\text{under 40}} - 0.5U_i \\ &\quad - 0.5(1-U_i)X_{i,\text{under 40}} + 1.5U_iX_{i,\text{under 40}} + 0.4U_iX_{i,\text{under 40}}X_{i,\text{white}} \\ &\quad + 0.5Z_i - 0.25U_iZ_i \\ g(\mathbf{x}_i, u_i, z_i) &= -3.1 + 0.5X_{i,\text{white}} - 1.5X_{i,\text{under 40}} + U_i + 0.5Z_i - 0.2U_iZ_i \end{aligned}$$

Figure 4.1 shows how the degree of selection bias (*ddc*) and the explanatory power of auxiliary variables $\mathbf{X}$ and U (measured as the R^2 of a simple regression of Y on $\mathbf{X}$ and U) vary with the amount of observation noise, γ introduced in simulations. When γ is close to 0, there is very little observation noise and selection

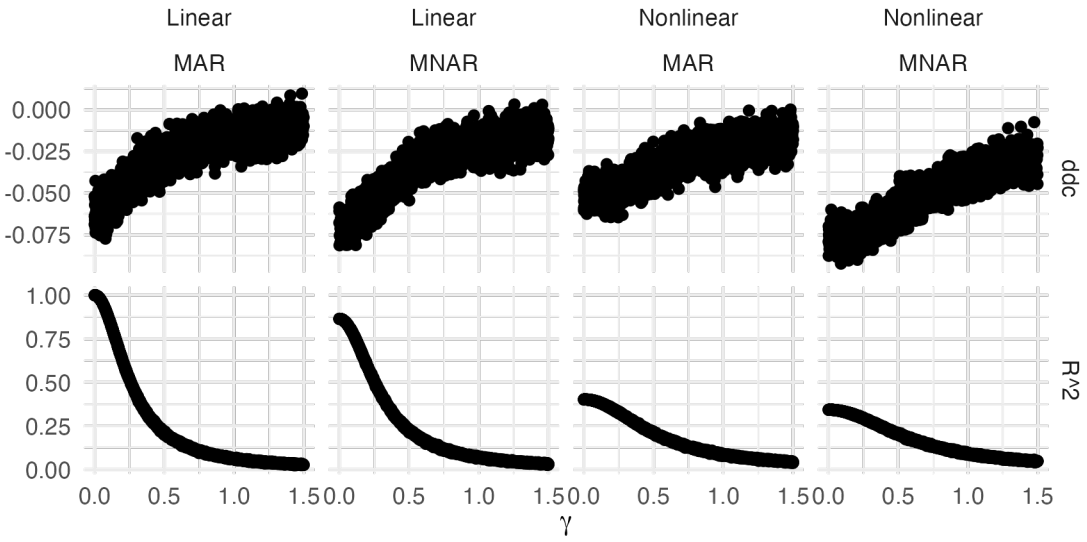


Figure 4.1: Variation in selection bias, *ddc*, and explanatory power of auxiliary covariates $\mathbf{X}$ and U, R^2, across simulation settings and degree of observation noise, γ.

bias and the explanatory power of auxiliary variables is maximized. As γ increases, both *ddc* and R^2 go to 0. In both the linear and nonlinear MNAR settings, R^2 decreases relative to the corresponding MAR setting

It is impossible to estimate "true" leverage in most practical settings, which requires estimation of $\mathbb{E}_{\mathbf{X}}[f(\mathbf{x},1) - f(\mathbf{x},0)]$ for all values of $\mathbf{x}$, as well as estimation of the population standard deviations σ_U and σ_U. In simulation settings, however, we are able to directly estimate these quantities, so are able to compare leverage estimated using each estimation method to "true" population leverage. In particular, we assess the performance of four methods for estimating leverage:

- **Regression**: Leverage is estimated as the standardized regression coefficient in a regression of Y on $\{\mathbf{X}, U\}$, $\hat{\lambda}_{\text{reg}} = \hat{\beta}^*_u$
- **Perturbation - fixed** δ: Leverage is estimated as $\hat{\lambda}_\delta = \frac{\bar{Y}_{\tilde{w}} - \bar{Y}_w}{\sigma_Y \times \delta_U}$ where $\delta_U = \frac{\mathbb{E}[\tilde{U}] - \mathbb{E}[U]}{\sigma_U}$. Here we set $\delta_U = \text{sign}(0.5 - \mathbb{E}[U])$, such that $\mathbb{E}[\tilde{U}] = \mathbb{E}[U] \pm \hat{\sigma}_U$. We use $\text{sign}(0.5 - \mathbb{E}[U])$ to set δ_U so that $\mathbb{E}(\tilde{U})$ is always closer than $\mathbb{E}(U)$ to 0.5, where uncertainty for a binary variable is maximized. This also ensures that we avoid $\mathbb{E}(\tilde{U})$ outside of $[0, 1]$.
- **Perturbation - drop** U: Leverage is estimated as $\lambda_d = \frac{\bar{Y}_{\tilde{w}} - \bar{Y}_w}{\sigma_Y \times \delta_U}$ where $\tilde{w}$ are estimated only using $\mathbf{X}$, no U.
- **Conditional Expectation**: Leverage is estimated as $\hat{\lambda}_{\text{ce}} = \frac{1}{n} \sum_{\mathbf{X}} n_{\mathbf{x}} (\bar{Y}_{\mathbf{x},1} - \bar{Y}_{\mathbf{x},0})$

4.3.2 Results

Figure 4.2 shows the results of the simulation, categorized by degree of selection bias as measured by the absolute value of the population *ddc*. In settings where the true $f(\mathbf{x}, u, z)$ is a simple linear function, there is little difference in performance between in the methods. The regression-based estimator of leverage tends to be lower-variance and have slightly lower median absolute error in low-*ddc* settings, however performs worse than other methods in settings with the most severe selection bias. This pattern is consistent across both MNAR and MAR settings.

The regression-based estimator starts to deteriorate in settings in which the true $f(\mathbf{x}, u, z)$ is not just a simple linear model and instead includes interaction terms, and absolute estimation error increases markedly as selection bias worsens. In these settings, regression-based estimation performs significantly worse than direct estimation of conditional expectation, or estimation by perturbation, methods that appear to be robust to assumptions about the functional form of f. Direct estimation of conditional expectation directly accounts for interactions between $\mathbf{X}$ and U as $f(\mathbf{x}, 1) - f(\mathbf{x}, 0)$ is estimated for each value of $\mathbf{x}$. Perturbation methods, however, do not explicitly account for any interactions as weighting is performed via raking to marginal distributions of each $\mathbf{X}$ and U, however if performs just as well as conditional expectation.

Estimation by conditional expectation is analogous to IPW using post-stratification and relies on the full joint distribution of $\mathbf{X}$ and U, whereas estimation by perturbation is analogous to (and directly uses) raking. Conditional expectation likely performs well here due to the small number of strata considered (8 total), and, as with post-stratification weighting, likely deteriorates as the set and granularity of auxiliary features increases. Therefore, in more complex settings, perturbation may outperform conditional expectation.

4.4 Benchmark uncertainty in the Axios-Ipsos Coronavirus Tracker

We use the Axios-Ipsos Coronavirus Tracker (Jackson et al., 2021) to explore using leverage to assess uncertainty in the population frame. The Axios-Ipsos Coronavirus Tracker fielded every 2-3 weeks from March 2020 through December 2022 and tracked public opinion related to COVID among American adults. In particular, starting January 8, 2021, the survey asked respondents whether they had received at least one dose of a COVID vaccine. Reaching a threshold of first-dose vaccine uptake was

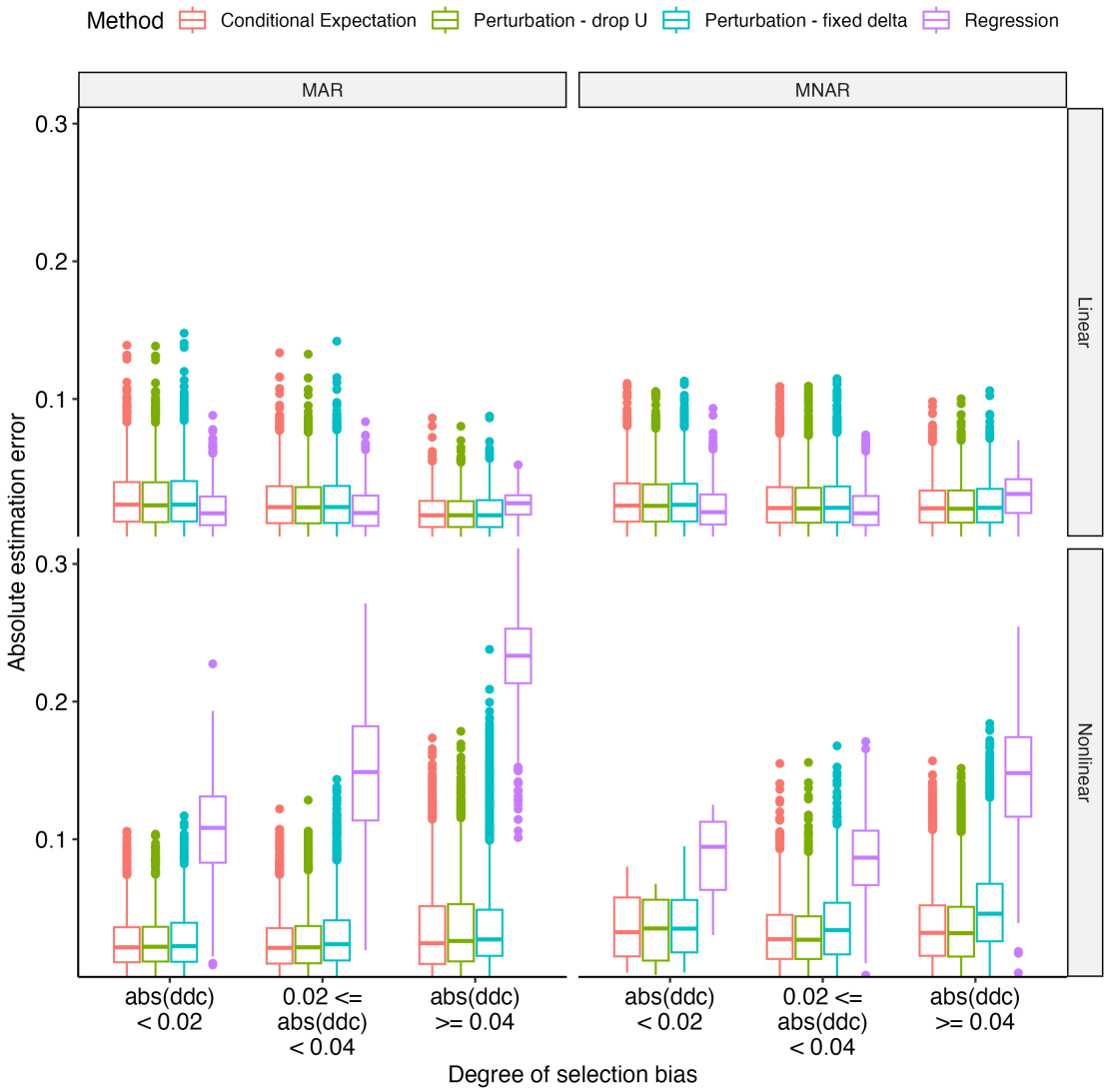


Figure 4.2: Simulation results showing absolute error for each estimation method by type of missingness, functional form of f, and the severity of selection bias. Error is measured relative to the true population leverage, $\mathbb{E}_{\mathbf{X}}[f(\mathbf{x}, 1) - f(\mathbf{x}, 0)]$.

widely viewed as a necessary benchmark for relaxing COVID restrictions, and as a result, was monitored vigorously by survey researchers, governments, and the public.

However, surveys tracking COVID vaccination rates varied widely in their estimates of first-dose uptake among American adults, and some differed drastically from CDC estimates, which collated actual numbers of doses administered from entities performing vaccinations (though it also suffered from bias and reporting delays) (Bradley et al., 2021). Axios-Ipsos' estimates tracked closely with the CDC, and was one of the only surveys to account for political partisanship of respondents. However, there is no single source of truth for the partisanship distribution of American adults, so Ipsos weights their Coronavirus tracker to

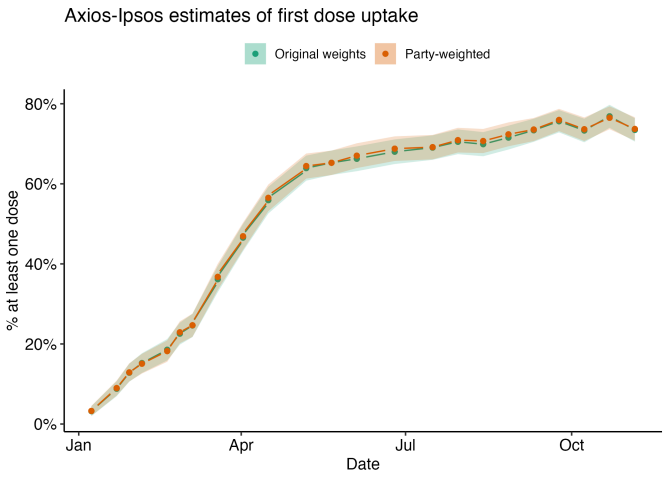


Figure 4.3: Estimates of first dose COVID vaccine uptake from Axios-Ipsos Coronavirus tracker with original weights ('Raw'), and reweighted to have consistent partisanship ('Party-weighted'). Shaded area shows 95% confidence intervals incorporating variance inflation from weighting.

partisanship estimates "from recent ABC News/Washington Post telephone polls." Although these estimates are treated as ground truth, they are in fact survey-based estimates and, even if they are unbiased, are uncertain.

In standard practice, this benchmark uncertainty is left unquantified, if not ignored entirely. However, we demonstrate here how leverage can be used to help *analytically* quantify the uncertainty in estimates of vaccine uptake that results from uncertainty in the underlying partisanship benchmark used in weighting. The exact partisanship targets that Ipsos uses are not clearly defined, and weighting on partisanship is only employed when the raw data falls outside of a certain (undisclosed) margin of the weighting targets. Therefore, we first re-weight the Axios-Ipsos data to have consistent partisanship margins as defined by the average of 4 ABC News/Washington Post national political polls that were fielded in 2021 (the same time period as the Axios-Ipsos data). Figure 4.3 shows estimates of first dose uptake with original weights and with consistent partisanship weights. Estimates of first dose uptake increase slightly with consistent party weighting, but are largely unchanged.

Partisanship as used by Ipsos and ABC News/Washington Post has 4 categories: Democrat, Republican, Independent, and Something else. We have defined leverage in the binary case, so here consider each category separately and calculate the leverage for each in each wave of the Axios/Ipsos Coronavirus Tracker using the perturbation method outlined in Section 4.2.

Estimates of leverage of partisanship on first-dose COVID vaccine uptake are shown in Figure 4.4. The leverage of partisanship on first dose COVID vaccine

uptake changes over time as vaccines become more widely available. At the start of the vaccine roll-out in January 2021, the leverage of any party affiliation on estimates of vaccine uptake is negligible. During this time, access to vaccines was highly regulated and limited to front-line workers and the most medically vulnerable populations. Expanding access to vaccines throughout the winter and spring of 2021 coincides with an increase in the magnitude of leverage of Democratic and Republican partisanship. Large positive leverage of Democratic partisanship indicates that an increase in the weight of Democratic respondents in a sample leads to higher weighted estimates of first-dose uptake. Conversely, large negative values of leverage for Republican partisanship indicate that increasing the weight of Republicans in a sample leads to lower estimates of first-dose uptake. Changes in partisanship leverage throughout 2021 reflect the impact of partisanship on American adults' vaccination status as logistical barriers to vaccination are eliminated and vaccination status is a matter of personal preference rather than access to vaccines.

Estimates of first-dose vaccine uptake will be the most sensitive to error in the survey estimates used to derive partisan weighting targets when the magnitude of partisan leverage is highest (May-July of 2021). We can use estimates of leverage to derive a population frame uncertainty margin of error (MoE) for first-dose vaccine uptake (as described in Section 4.2.4 and in Appendix 4.8). We set m equal to $n_{\text{eff}} = 3{,}917/1.28 = 3{,}053$ which is the effective sample size across the 4 ABC News/Washington Post polls used to estimate partisanship rates. Figure 4.5 shows the distribution of population uncertainty MoEs for each partisanship category, with median MoE size as dotted lines. Democrat and Republican partisanship weighting targets have similar contributions to error across survey waves. In the median wave, a ±1pp error in the population proportion of Democrats (or Republicans) used as a weighting target leads to an *additional* ±0.5pp of error in estimates of first-dose vaccine uptake, beyond the uncertainty from sampling and weighting. Although this error may seem small, in 2020, the margin of victory for Biden in WI was 0.6pp, so error of this magnitude can lead to incorrectly forecasting the winner of a US Presidential election. An extension of Bradley et al. (2021) shown in Appendix Figure 4.6 shows that Axios-Ipsos tracks the CDC benchmark quite well from January to March 2021, but beginning in April, standard confidence intervals do not include the CDC's estimates of first-dose uptake as consistently. Including population frame uncertainty from partisanship may help improve coverage.

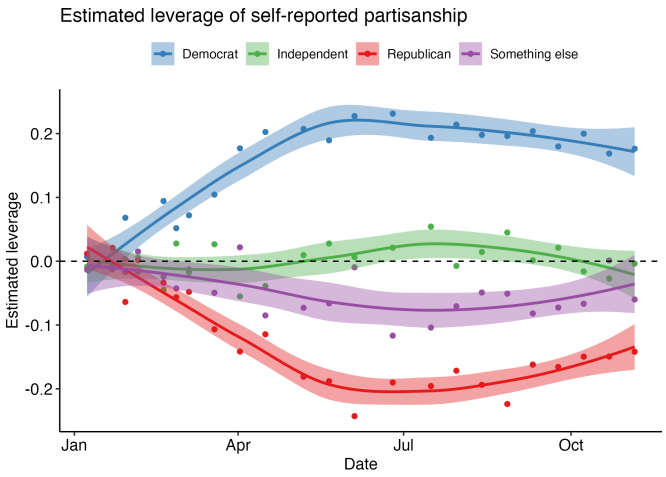


Figure 4.4: Leverage of party affiliation in Axios-Ipsos Coronavirus tracker surveys on estimates of first dose COVID vaccine uptake. Leverage is estimated using perturbation.

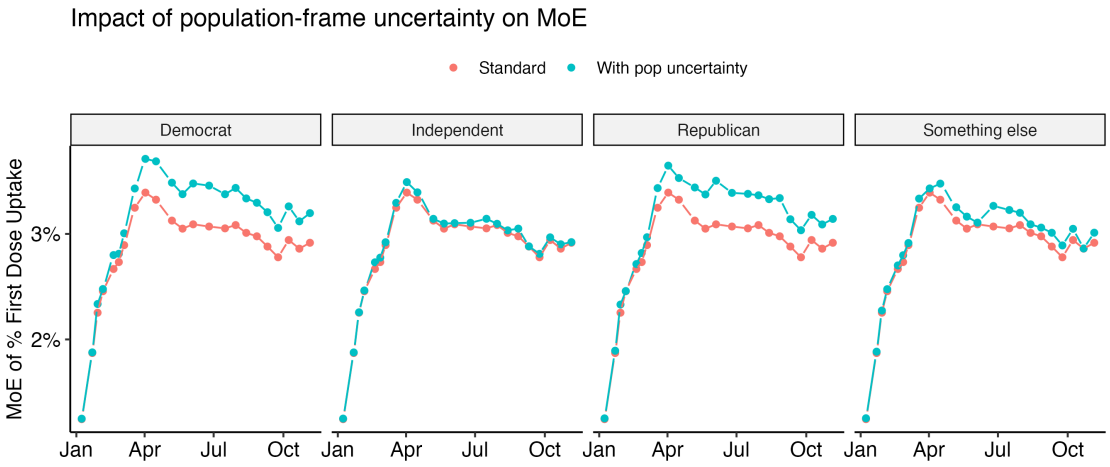


Figure 4.5: The width of 95% MoEs for first-dose vaccine uptake estimated using standard methods ("Standard"), and incorporating population frame uncertainty about the true rate of partisanship affiliation among American adults.

4.5 Discussion

Leverage is a novel method for performing sensitivity analysis in survey weighting. Leverage can be estimated using any IPW method, so is easily incorporated into existing survey workflows. In comparison to other survey weighting sensitivity techniques, leverage tackles a different, and often ignored, source of uncertainty in surveys – that from imperfect population frame targets. We provide tools here for using leverage to incorporate such uncertainty into traditional survey MoEs, and in our application to the Axios-Ipsos Coronavirus tracker, demonstrate how tracking

leverage over time or across similar surveys can provide a new dimension of insights into the survey data generating process. Beyond its use as a tool for sensitivity analysis, we hope that this work provides survey researchers with a simple tool for quantifying an under-appreciated source of error.

A critical shortcoming of leverage as defined here is its assumption that covariates are binary. We show in applications to the Axios-Ipsos Coronavirus Tracker how leverage can be adapted to fit multi-category covariates, but future research is needed into how best to implement leverage for multi-categorical, continuous, and interacted covariates.

Appendix

4.6 Appendix: Existing Methods for selecting X

There are a variety of existing methods for selecting auxiliary variables for nonresponse adjustment. The goal of nonresponse adjustment is to induce conditional independence between (to *d-separate*) an outcome Y and response R by conditioning on a set of auxiliary variables X (Pearl, 1995a). Most methods seek to address this by aiming to select the set $\mathbf{X}$ that best predicts each R and Y.

Geuzinge et al. (2000) suggest selecting variables that maximize the quantity $\hat{\rho}_{Y,X} \times \rho_{R,X}$, where $\hat{\rho}_{Y,X}$ is the sample correlation between Y and a proposed X and $\rho_{R,X}$ is the population correlation between a proposed auxiliary variable X and the recording indicator. Note that this method requires that $\mathbf{X}$ is observed for the entire population, and assumes that $\hat{\rho}_{Y,X}$ is an unbiased estimate of population $\rho_{Y,X}$.

Bethlehem and Schouten (2004) expand on this approach to consider *sets* of auxiliary variables instead of each independently. They propose a step-wise procedure for selecting the subset of $\mathbf{X}$ that minimizes the maximum absolute bias of the generalized regression estimator for Y (Deville and Särndal, 1992). They find that minimizing the maximum absolute bias of $\bar{Y}$ is equivalent to minimizing

$$W(X) = \sqrt{1 - \rho^2_{\pi, X\beta_{X,Y}}} \sqrt{1 - \rho^2_{X\beta_{X,Y},Y}} \tag{4.36}$$

where $\pi = P(R = 1)$, $\beta_{X,Y}$ is the vector of coefficients from regressing Y on the set of selected auxiliary variables $\mathbf{X}$, and ρ denotes correlation. Estimating $\rho_{X\beta_{X,Y},Y}$ is not possible because Y is only observed in the sample, so instead it is estimated by the sample correlation $\hat{\rho}_{X\beta_{X,Y},Y}$, producing the modified metric W^*. Furthermore, while we observe the response indicators R_i for the entire population, we do not observe the true underlying π_i, and therefore must estimate it using available covariates X. Intuitively, the bias of the generalized regression estimator for Y will decrease as the explanatory power of the set of covariates used in adjustment increases.

Caughey and Hartman (2017) iterates on the W^* metric of Bethlehem and Schouten (2004) and instead of using the proposed step-wise approach, leverages the LASSO for selecting a model for each Y and R, and weighting on the joint set of significant features. In practice, the method is highly sensitive to noise in the model specification, and often results in a selected set of auxiliary variables that is too large for standard weighting techniques (Bradley and Nichols, 2022).

So far, all the methods discussed have been outcome-dependent, however Särndal (2008) introduces an outcome-agnostic method for selecting a set of auxiliary variables. Auxiliary variables are selected to maximize

$$Q^2 = \text{Var}(1/\hat{\pi}_i), \tag{4.37}$$

where $\hat{\pi}$ is the predicted probability of response, defined by some function of selected $\mathbf{X}$. Q^2 will increase as the predictive power of the set of auxiliary variables increases, and thus $\hat{\pi}_i$ better predicts true response probabilities. If π_i were known, we would be able to perfectly recover from selection bias, but as we cannot, Q^2 measures our ability to predict it from $\mathbf{X}$.

Hartman et al. (2021) forgo the need for variable selection at all by approaching the problem as one of dimensionality reduction. The **kpop** approach uses standard calibration to derive weights that ensure that the first r eigenvectors of the sample kernel matrix $K = k(x_i, x_j)$ match those of the population. Not only does this approach address the problem of variable selection, but also does not require the assumption that the probability of response is a simple function of independent covariates and accounts for important higher-order interactions of X.

Underlying these methods is the assumption that selection bias does not affect estimates of associations, i.e. $\hat{\beta}_{X,Y}$ or $\rho^*_{X\beta_{X,Y},Y}$, which we know can fail in the presence of collider bias (Munafò et al., 2018). Furthermore, all methods discussed thus far require that values of all potential auxiliary variables are observed for the entire population, not only for sampled units. Population data can be extremely costly to come by, even in aggregate, and thus this limitation of existing methods can make their use prohibitively resource-intensive in practice. The reliance on individual-level population data also limits the extent to which these methods can be used for exploratory analysis and monitoring of the influence of particular auxiliary variables as patterns of nonresponse evolve. Leverage seeks to address these shortcomings.

This method is related to the weighting sensitivity analysis method introduced in Hartman and Huang (2024), however focuses on the bias that results from the exclusion of a key auxiliary variable, whereas here we consider the bias driven by error in weighting targets for an auxiliary variable.

4.7 Appendix: Data Defect Correlation *ddc*

Meng (2018) decomposes the total error in the estimate of a population mean, $\bar{Y}_n - \bar{Y}_N$, into three interpretable quantities:

$$\bar{Y}_n - \bar{Y}_N = \hat{\rho}_{Y,R} \times \sqrt{\frac{N-n}{n}} \times \sigma_Y \tag{4.38}$$

The first component of error is $\rho_{Y,R}$, the correlation between the outcome of interest Y and the recording indicator R. This is the *data defect correlation* and measures the sign and degree of selection bias. The second component, $\sqrt{\frac{N-n}{n}}$, measures the impact of data *quantity* on error, where N is the population size and n is the sample size, which goes to 0 as the fraction of the population observed n/N approaches 1. The last quantity, σ_Y, the population standard deviation of Y, measures population heterogeneity of the outcome.

An additional quantity, the *data defect index* or *ddi* is defined as the square of *ddc*, $\rho^2_{Y,R}$, and is a unified index of data quality. *ddi* ranges from 0 to 1, with higher values indicating lower data quality.

Estimating *ddi* directly requires that both Y and R are observed for the entire population, which is generally impossible. Estimating *ddi* using modeled or imputed values of Y for population units can suffer severely from collider bias (as we will demonstrate in simulations). Instead, we can estimate *ddi* in practice if we observe a population benchmark for Y, as demonstrated in Meng (2018) and Bradley et al. (2021). This is still rare in most settings, but it more feasible than observing individual-level outcome data for the entire population.

In the simulation setting, N, n, and σ_Y are fixed, and Y is known for the population. This means that we can calculate $\rho_{Y,R}$ directly, and that error is directly correlated with data quality in unweighted samples. However, an alternate form of Meng's identity allows for the incorporation of survey weights, which not only impact data quality (hopefully improving *ddi*), but also generally reduce the effective sample size. Using *ddi* instead of absolute or squared error allows us to account for the impact on estimator variance from weighting.

4.8 Appendix: Derivation of population frame uncertainty interval

Assume we have not just an estimate of $\mathbb{E}(U)$, but also some estimate of uncertainty for it. Then, we can use leverage to estimate how uncertainty in $\mathbb{E}(U)$ translates to uncertainty in $\bar{Y}_w$.

Say that we are interested in estimating $\bar{Y}_N$ using a survey of size n, and use IPW with auxiliary features $\{\mathbf{X}, U\}$ to adjust for observed selection bias. Our population target $\mathbb{E}(U)$ is estimated from another survey (e.g. a large government-run survey like the American Community Survey) of size m. We construct a standard α_u-level confidence interval for the true $\mathbb{E}(U)$ based on the sampling distribution of U:

$$\bar{U}_m \pm \frac{\sigma_U}{\sqrt{m_{\text{eff}}}} \times Z_{1-\alpha_u/2}$$

where m_{eff} is the effective sample size from weighting, given by m/deff.

We can then use leverage to estimate $\bar{Y}_w$ if we weighted to population targets for U defined by the bounds of the confidence interval for $\mathbb{E}(U)$, which we will denote $\bar{U}^{lb}$ and $\bar{U}^{ub}$ for the lower bound and upper bound, respectively.

Recall that leverage is defined as $\hat{\lambda} = \frac{\bar{Y}_{\tilde{w}} - \bar{Y}_w}{\sigma_Y \times \delta_U}$, where $\delta_U = \frac{\mathbb{E}[\tilde{U}] - \mathbb{E}[U]}{\sigma_U}$, i.e. it is the number of population standard deviations of U that $\mathbb{E}(\tilde{U})$ falls away from $\mathbb{E}(U)$. We can rearrange the definition of leverage to estimate $\bar{Y}_{\tilde{w}}$ given a certain $\mathbb{E}(\tilde{U})$:

$$\begin{aligned}\bar{Y}_{\tilde{w}} &= \bar{Y}_w + \hat{\lambda} \times \sigma_Y \times \delta_U \\ &= \bar{Y}_w + \hat{\lambda} \times \sigma_Y \times \frac{\mathbb{E}[\tilde{U}] - \mathbb{E}[U]}{\sigma_U}\end{aligned}$$

$\mathbb{E}_{lb}(U)$ and $\mathbb{E}_{ub}(U)$ are each $\frac{\sigma_U}{\sqrt{m_{eff}}} \times Z_{1-\alpha_u/2}$ from $\mathbb{E}(U)$, by definition of a sampling confidence interval. Therefore based only on normal sampling variation in U, we would expect $\bar{Y}_w$ to fall within $\left(\bar{Y}_{\tilde{w}_{lb}}, \bar{Y}_{\tilde{w}_{ub}}\right)$ estimated as

$$\begin{aligned}\left(\bar{Y}_{\tilde{w}_{lb}}, \bar{Y}_{\tilde{w}_{ub}}\right) &= \bar{Y}_w \pm |\hat{\lambda}| \times \sigma_Y \times \frac{\frac{\sigma_U}{\sqrt{m_{eff}}} \times Z_{1-\alpha_u/2}}{\sigma_U} \\ &= \bar{Y}_w \pm |\hat{\lambda}| \times \sigma_Y \times \frac{Z_{1-\alpha_u/2}}{\sqrt{m_{\text{eff}}}}\end{aligned}$$

where we use the absolute value of λ_U to ensure that $\bar{Y}_{\tilde{w}_{lb}} < \bar{Y}_{\tilde{w}_{ub}}$.

This uncertainty interval does not account for variance from sampling or weighting in the survey we are using to estimate $\bar{Y}_w$, it only accounts for uncertainty in $\mathbb{E}(U)$. Therefore, we can account for this additional uncertainty by specifying confidence intervals for each $\bar{Y}_{\tilde{w}_{lb}}$ and $\bar{Y}_{\tilde{w}_{ub}}$, each with width $\frac{\sigma_Y}{\sqrt{n_{\text{eff}}}} \times Z_{1-\alpha_Y/2}$. Therefore, the full uncertainty interval for $\bar{Y}_w$ that accounts for normal sampling variability of Y, variance inflation from weighing, and uncertainty from sampling variation in U is given by

$$\bar{Y}_w \pm \left(|\hat{\lambda}| \times \frac{\sigma_Y}{\sqrt{m_{\text{eff}}}} \times Z_{1-\alpha_u/2} + \frac{\sigma_Y}{\sqrt{n_{\text{eff}}}} \times Z_{1-\alpha_y/2}\right) \tag{4.39}$$

Note that α_u and α_y do not have to be the same. This should not be interpreted strictly as a confidence interval because it attempts to combine two different sources of uncertainty, but rather as a type of sensitivity analysis. Also, this interval only accounts for population frame uncertainty in a single binary U. If multiple population weighting targets are estimated and not known for certain, then this interval likely still underestimates the true uncertainty in $\bar{Y}_w$.

4.9 Appendix: Extension of Bradley et al. (2021)

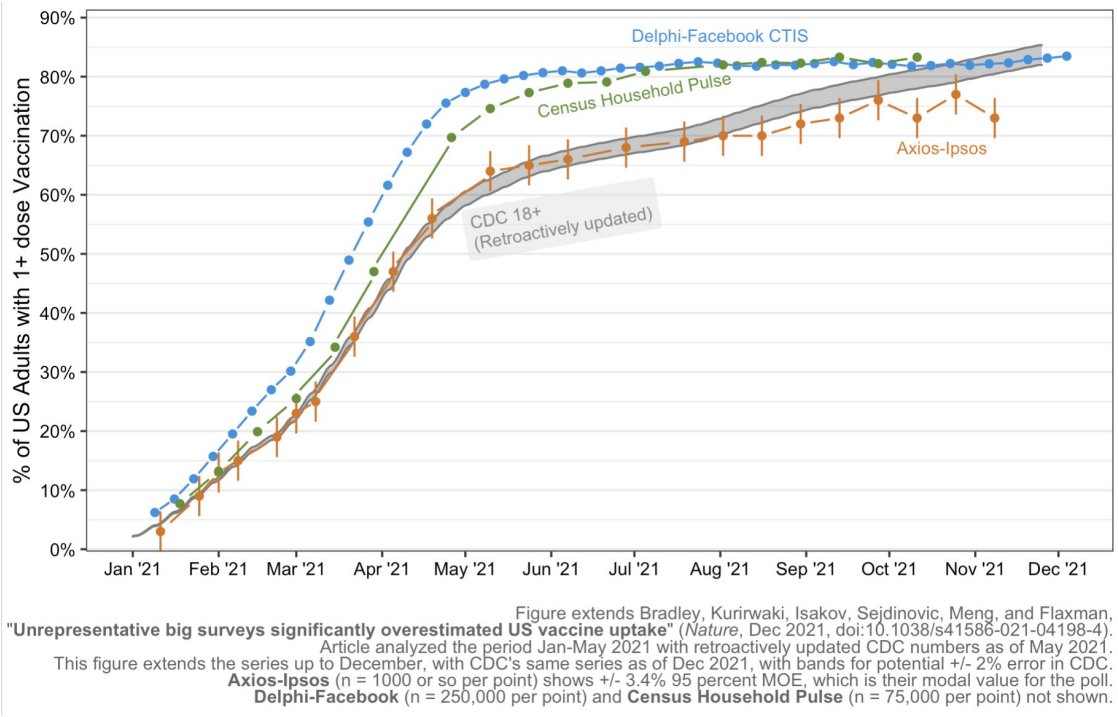


Figure 4.6: An extension of Figure 2 from Bradley et al. (2021) that uses surveys over a large time frame, and incorporates updated CDC benchmark data.

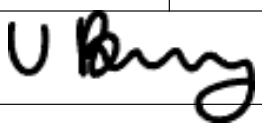
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Leverage of weighting auxiliary variables
Publication Status	☐Published ☐ Accepted for Publication ☐Submitted for Publication ☒Unpublished and unsubmitted work written in a manuscript style
Publication Details	**Bradley V., Schwenzfeier, M., and Sejdinovic, D. Leverage of weighting auxiliary variables.**

Student Confirmation

Student Name:	Valerie C Bradley		
Contribution to the Paper	Jointly developed the idea with M.S., jointly developed the theory with D.S., performed simulations, wrote manuscript		
Signature 		Date	Jan 5, 2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: **Prof Dino Sejdinovic**		
Signature 	Date	7 January 2024

This completed form should be included in the thesis, at the end of the relevant chapter.

5

Methods for selection bias adjustment in the UK Biobank neurological imaging data

Abstract

We review the structural causal model (SCM) literature, including the conditions under which is it possible to recover from selection bias, and the most common methods for adjusting for selection bias, including inverse probability weighting (IPW) and regression methods. We introduce a new adjustment method that uses Bayesian Additive Regression Trees (BART) and raking, designed to leverage the advantages of each IPW and regression methods. These methods are first tested in simulation studies where true selection bias is known. They are then applied to data from the UK Biobank imaging cohort, a subset of the UK Biobank, a national prospective health study of half a million participants between the ages of 40 and 69, which is known to suffer from selection bias. The 2016 Health Survey for England is used to define the target population.

5.1 Introduction

The UK Biobank is a national prospective study of half a million participants between the ages of 40 and 69 at the time of recruitment between 2006 and 2010. The UK Biobank was established to examine relationships between exposures and common health-related outcomes that affect aging populations like cancer, heart disease, diabetes and dementia (Sudlow et al., 2015). Though the study design took steps to maximize the generalizability of the UK Biobank cohort, recruiting enough participants for analysis of complex exposure-outcome relationships was of greater concern (Sudlow et al., 2015). As a result, Fry et al. (2017) describes how the cohort suffers from "healthy volunteer" bias, in that participants exhibit lower rates of smoking, obesity and daily alcohol consumption than the target population of UK adults. Strikingly, Fry et al. (2017) note that "all-cause mortality is approximately half that of the UK population as a whole, and total cancer incidence rates are approximately 10%-20% lower." The authors conclude that the UK Biobank is not representative of the population, but that the cohort can still be used to draw valid, generalizable conclusions about exposure-disease relationships, which is the primary concern of the study.

As of 2021, the UK Biobank was in the process of recruiting a subset of the total 500,000 UK Biobank participants to undergo additional assessments as part of the world's largest ever multi-modal imaging study. Since recruitment began in 2016, over 30,000 participants have completed the additional imaging screenings, with the goal of completing 100,000 in total. Littlejohns et al. (2020) begin to examine the representativeness, or lack thereof, of the UK Biobank imaging cohort and finds that participants have lower rates of smoking and obesity than the larger UK Biobank cohort. This suggests that the imaging cohort suffers from even more "healthy volunteer" selection bias that the full UK Biobank cohort, possibly due to the high burden associated with undergoing extensive imaging assessments. However, Littlejohns et al. (2020) asserts that drawing inferences about exposure-outcome associations are still valid.

Although selection bias is most commonly acknowledged to impact estimates of prevalence, a form of selection bias known as "collider bias" can also significantly bias estimates of associations (Munafò et al., 2018). This bias may undermine the validity of drawing conclusions about associations from UK Biobank data in general, and the UK Biobank imaging cohort perhaps to an even greater degree. This paper seeks to explore methods for ameliorating selection bias, with a particular focus on collider bias in the UK Biobank imaging cohort.

In addition to extensive survey research literature on selection bias, there are a number of papers that examine selection bias in clinical settings. LeWinn et al. (2017) use raking, one of the methods we consider in this paper, to adjust a sample of 1,162 structural brain images from a community-based sample of children, and find altered associations between age and brain structure compared to unadjusted associations. Other research has explored methods for adjusting for selection bias in longitudinal cohort studies (Nohr and Liew, 18), in epidemiology (Infante-Rivard and Cusson, 2018), and more generally across a variety of application areas (Bareinboim et al., 2014).

The primary focus of this paper is validating a mix of existing and novel methods for adjusting selection-biased data so that they may still be used to draw generalizable inferences. We seek to ground these practical methods in the robust statistical framework of structural causal models (SCMs) pioneered by Judea Pearl. SCMs provide a language for expressing the theoretical conditions under which causal, associative, and prevalence inferences may be drawn from selection biased-data. These conditions motivate our approach to validation and provide theoretical justification for the advantages of our novel adjustment procedure.

The paper is outlined as follows. Section 5.2 serves as a primer on structural causal models, admissible sets, and inverse probability weighting (IPW) procedures, and Section 5.3 presents our novel IPW procedure. Section 5.4 describes the simulation studies used to validate IPW methods, and Section 5.5 gives an in-depth overview of the UK Biobank data, including the key outcomes we will use to evaluate adjustment procedures, and the Health Survey for England (HSE) and UK Census data used for adjustment. Section 5.6 presents results quantifying the bias that exists in the UK Biobank imaging cohort, results from simulation studies comparing adjustment procedures under consideration and results from estimation of key UK Biobank outcomes using adjustment procedures. Section 5.7 discusses the relative advantages of adjustment procedures and explores areas for future research.

5.2 Methods: Background

5.2.1 Selection bias

Selection bias is a problem experienced by almost any researcher working with real data. Heckman (1979) describes selection bias as the bias that "results from using non-randomly selected samples to estimate behavioral relationships." In practice this often presents as subjects who have differentially self-selected into, been excluded from, or withdrawn from a study, giving rise to effects like volunteer

bias, non-response bias and differential loss-to-follow up. Selection bias may also result from decisions made by the analyst when preparing data for analysis. For example, a researcher may choose to exclude subjects with item nonresponse to an outcome of interest, even though the subject participated in the rest of the study.

Hernán et al. (2004) presents a unified, structural approach to selection bias by identifying a common causal structure - collider bias - underlying biases usually placed under the umbrella of "selection bias." A collider is the common effect of two exposure variables, and conditioning on the collider can induce spurious correlation between the two, perhaps truly unrelated, exposures. Hernan argues that selection biases are commonly the result of conditioning on a collider for which one cause is the exposure of interest (or a cause of the exposure) and another is the outcome of interest (or a cause of the outcome of interest).

Selection bias can occur without a collider (e.g. prevalence studies may suffer from selection bias), and similarly collider bias can occur outside of the scope of selection bias (e.g. analysis that conditions on confounding variable that is not related to selection). However, the structure of collider bias is useful for analyzing the impact of common forms of selection bias in studies like the UK Biobank that focus mainly on examining associations.

To illustrate the impact of collider bias, consider an example based on the UK Biobank. We observe two variables: retirement status (1 if retired, 0 if otherwise), and travel time in minutes from home to the closest UK Biobank assessment center. Retirees have more time to participate in the UK Biobank than those who may have to take time off of work to attend an assessment. Let us also assume that participation rates decline monotonically as travel time increases. We only observe people who participate in the study (the collider), whose decision to participate was influenced by their travel time and retiree status (common causes). Therefore, knowing that a participant is not retired provides information about their travel time to an assessment center (it is likely shorter than average). Conditioning on participation in the UK Biobank alters the observed association between travel time and retiree status - this is collider bias. The direction and magnitude of collider bias depends on the true underlying relationship between common causes, but will be present regardless of the true relationship between exposures (Nguyen et al., 2019).

5.2.2 How concerned should we be about selection bias?

Though the example above is trivial, it seems somewhat clear that selection bias should concern any researcher interested in drawing valid associative conclusions from unrepresentative data. However, there are those who argue otherwise. For example, Fry et al. (2017) claims that despite evidence of selection bias in the UK Biobank, "valid assessment of exposure-disease relationships may be widely generalizable and does not require participants to be representative of the population at large."

Rothman et al. (2013) articulates a similar argument in a letter to the *International Journal of Epidemiology*. They argue that there is a difference between *scientific* inference, which seeks to understand the underlying causal mechanisms of nature, and *statistical* inference, which seeks to describe a population based on a sample from that population. They claim that for *scientific* inference, representativeness is not necessary, and should in some cases be actively avoided, if it means a study is better able to understand the mechanical pathway by which an exposure effects an outcome. Rothman et al. (2013) gives an example of a study designed to test the efficacy of a drug - it might be advantageous to sample equal numbers of participants from each of 3 age brackets (e.g. 25-35, 35-45, 45-55) in order to make equally-confident statements about the drug's efficacy for each age bucket, even if it means that the age distribution of the study does not reflect that of the larger population. This is to say that a study with such a design might have higher *internal* validity than if it had been perfectly representative. We observe this type of design in the UK Biobank - the study was intended to examine diseases of ageing, so recruitment was limited to people between the ages of 40 and 70. The UK Biobank also aims to allow researchers to examine diseases of ageing within small subgroups of the population, so the design prioritized scale over representativeness of participants. Both Rothman et al. (2013) and the UK Biobank acknowledge that given such design choices, generalizing inferences to under-observed populations may be difficult, or even impossible, but, given the goals of the UK Biobank, this seems like a logical trade-off.

However, in arguing that unrepresentative studies aimed at understanding associations or causation need not be concerned about selection bias, Rothman et al. (2013) and Fry et al. (2017) overlook the problem presented by selection bias that takes the form of collider bias. Selection bias does not just impact the *external* validity of a study, but, when it presents as collider bias, can actually impact the *internal* validity of a study by biasing estimates of associations. As we noted, the UK Biobank is decidedly, and by design, unrepresentative. While the intentional unrepresentativeness with respect to age may not present a problem given the goals

of the study, the UK Biobank is also unrepresentative of the UK population in the sense that participants are less likely than the UK adult population to have smoked or to be obese (Fry et al., 2017). One possible explanation for this discrepancy is that those who are more aware of their health are more likely to participate in a large health-related study and are also less likely to ever have smoked. Fry et al. (2017) also found that women were more likely to participate in the UK Biobank than were men. Although there may be some true association between gender and consciousness of one's health, the estimate of this association from the UK Biobank data likely suffers from collider bias.

Researchers face many trade-offs when designing studies, and loss of representativeness in favor of a focus on other priorities may be wise. However, selection bias can impact study outcomes even when they are not focused on generalizing to other populations. Selection bias, particularly when it presents as collider bias, should be considered a risk to both the internal and external validity of a study. Luckily, there exists a large literature on methods for ameliorating the problem of selection bias in unrepresentative studies, for estimates of both prevalence and causal effects.

5.2.3 Structural causal models

In an enormous study like the UK Biobank, there exists a complex dependence structure between genetic, demographic, socioeconomic, and health factors. In order to identify, and ideally adjust for, selection bias it is necessary to think carefully about the relationships between these factors. In this section, we will introduce structural causal models (SCMs), or graphical models for encoding complicated conditional dependence structures like these, as pioneered in Pearl (1995a). Though this paper does not aim to make causal claims based on the UK Biobank, SCMs provide a framework for taking a principled and rigorous approach to selection bias in complex systems, even outside of the context of causal inference.

Judea Pearl has been a pioneer in the world of causal inference research. He was the first to demonstrate the use of directed acyclic graphs (DAGs) as "a mathematical language for integrating statistical and subject-matter information," specifically information about dependence structures (Pearl, 1995a). Pearl has gone on to use DAGs to define conditions under which causal inferences can be drawn from non-experimental data, even in the presence of potential confounding factors (Bareinboim et al., 2014). This SCM framework for causal inference is parallel and complementary to the *potential outcomes* framework pioneered by the other "father of causal inference," Donald Rubin (Bareinboim and Pearl, 2016). For simplicity, we will present the major concepts of the SCM literature as originally

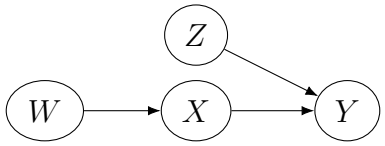


Figure 5.1: A simple structural causal model. Graph G with nodes X, Y and W.

conceived by Pearl in terms of discrete, non-parametric variables, however Pearl (1995b) extends the work to continuous and parametric cases.

Consider the directed graph G in Figure 5.1. G contains nodes X, Y and W which represent "physical mechanisms." The edges between nodes represent a direct causal pathway between two quantities, and the direction of the edge indicates the direction of the influence. X and Z are said to be the *parents* of Y because there are directed edges from X and Z to Y. Y is said to be the *child* of X and Y.

A lack of an edge between two nodes implies independence between those quantities. The edges between nodes represent *deterministic functions*, not just conditional dependence (though we could also express it as such), such that changes in X are due entirely to changes in its parents, denoted $pa_x = \{W\}$, or random changes in exogenous factors ϵ:

$$X = f(pa_x, \epsilon_x)$$

We can use such equations to describe all of the dependencies encoded in a graph. For example, graph G in Figure 5.1 can be described with the following equations:

$$\begin{aligned} Z = f_Z(\epsilon_Z), \quad W = f_W(\epsilon_W) \\ X = f_X(W, \epsilon_X), \quad Y = f_Y(Z, X, \epsilon_Y) \end{aligned}$$

Pearl (1995a) develops a "do-calculus," or a set of inference rules that can be used to translate the structural dependencies encoded in G into standard conditional probability statements. "Do calculus" uses a "do-operator" $do(X = x)$ to denote when a quantity is set to a particular value through experimental intervention rather than as a function of its parents, $X = f(pa_x, \epsilon_x)$, while all other mechanisms in the graph are unaffected. For example, in a graph with nodes X_i, $i = 1, \ldots, n$, we can express the joint distribution of the system as

$$p(X_1, \ldots, X_n) = \prod_i p(X_i | pa_{x_i}) \tag{5.1}$$

Referring again to Figure 5.1, the joint distribution of graph G would be:

$$p(Y, X, Z, W) = p(Y|X, Z)p(X|W)p(Z)p(W) \tag{5.2}$$

If we introduce a simple experimental intervention in which X is set to x, written $do(X = x)$, we would alter the dynamics of the system by removing the influence on X of its parent nodes pa_x or exogenous factors ϵ_x. We can describe the *causal effect* of this experimental intervention by updating the joint distribution to be conditional on $do(X = x)$:

$$p(Y, Z, W|do(x)) = \begin{cases} p(Y, X, Z, W)/p(X = x|W) & \text{if } X = x \\ 0 & \text{if } X \neq x \end{cases} \tag{5.3}$$

The equation above describes the impact of $do(x)$ on the entire system, but often we are interested in the causal effect on a single outcome Y, or $p(Y|do(x))$. We can isolate this quantity by integrating out the dependency on Z and W:

$$p(Y|do(x)) = \sum_z \sum_w p(Y, Z, W|do(x)) = \sum_z \sum_w p(Y|Z, W, do(x))P(Z)P(W) \tag{5.4}$$

We will revisit this equation in Section 5.2.5 to explain the conditions under which it is possible to identify causal effects even in the presence of selection bias.

Identifying conditions necessary for recovering from selection bias using SCMs also relies on the concept of *d-separation*. Nodes A and B are said to be *d-separated* by a set of nodes C if C blocks every path between nodes A and B (Pearl, 1995a). If A and B are d-separated by C then they are conditionally independent given C, $A \perp\!\!\!\perp B|C$. For example, in graph G, we could say that X d-separates W and Y, or $W \perp\!\!\!\perp Y|X$.

All of the concepts just introduced can be extended to the case where $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are disjoint sets of nodes instead of individual nodes (Pearl, 1995a), and to non-discrete state spaces. This includes showing that computing the effect of a continuous exposure variable X, where $X = g(Z)$ or $X = x$ with probability $p(x|z)$, is equivalent to computing $p(y|do(x))$.

In addition to encoding direct causal relationships, causal DAGs reveal structures that drive non-causal statistical associations. There are three types of structures that produce associations between variables (Hernán et al., 2004):

- **Cause and effect**: A direct causal path between exposure X and outcome Y
- **Common causes**: X and Y can be associated when they share a common cause even if they are not connected by a direct path. A change in the common cause drives changes in both X and Y.

- **Common effects**: When two exposures have a common effect, conditioning on the common effect induces an association between the two exposures, whether or not they are unconditionally related. The common effect is known as a *collider*.

Direct cause and effect relationships are often obscured by statistical associations induced by common causes and common effects. Causal DAGs and Pearl's do-calculus provide a language for expressing the complicated dependence structures that result in such non-causal relationships. In the next section, we will provide more detail about how these structures complicate causal inference and give strategies for overcoming some of these cases.

5.2.4 Selection and confounding bias in SCMs

Confounding bias and selection bias are the two main forms of bias that affect a researcher's ability to make robust, generalizable statistical inferences from observational data. These biases are driven either by common causes or common effects between exposure and outcome.

Confounding bias occurs when relationships between two variables are influenced by other unobserved quantities, or confounders. In causal inference, random controlled trials (RCTs) are considered the "gold standard" because random treatment assignment controls for all possible confounders, thus making it possible to isolate treatment as the reason for differences in outcomes between groups. In observational data, treatment is not randomly assigned and causal pathways are often confounded by other quantities. More formally, *confounding bias* occurs when the effect on Y of a treatment X, $\mathbb{P}(y|do(x))$, is not equal to the conditional probability $\mathbb{P}(y|x)$ due to the existence of a "back-door path" from X to Y. Figure 5.2a shows a graphical model that suffers from confounding bias. The dotted line represents an path between X and Y through unobserved (confounding) factors; this path is considered "open."

Pearl (1995a) derives conditions for when causal effects are identifiable even in the presence of confounding bias. One such condition is the *back-door criterion* which says that the causal effect $\mathbb{P}(y|do(x))$ is identifiable if there is a set of observed nodes $\mathbf{Z}$, disjoint from X and Y, that blocks every indirect (back-door) path from X to Y. Figure 5.2b shows a setting similar to that of 5.2a, but in this case, the backdoor path from $X \to Y$ is "blocked" by observed covariates represented by the set of nodes $\mathbf{Z}$. $\mathbb{P}(y|do(x))$ is identifiable in 5.2b but not in 5.2a.

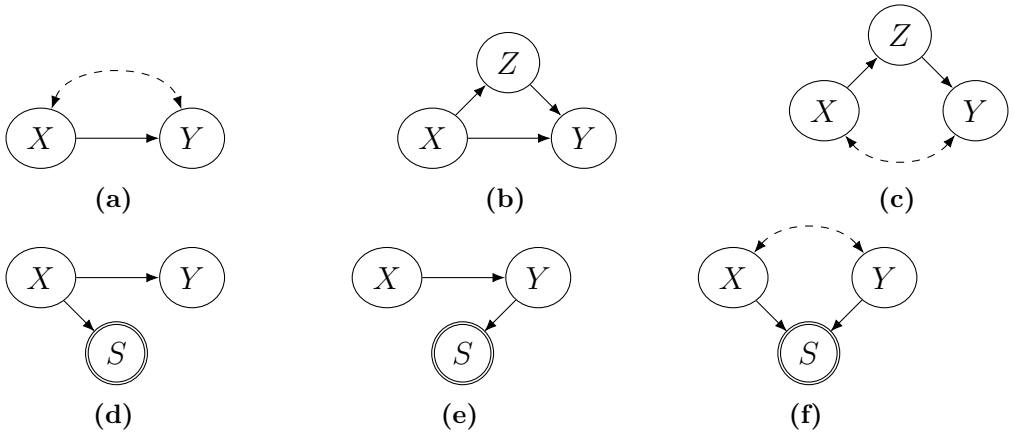


Figure 5.2: Simple examples of SCMs with confounding or selection bias. (a) a case of confounding where the causal effect is unidentifiable (b) confounding bias that is blocked by Z, satisfying the back-door criterion (c) a case in which there is confounding between X and Y, but Z satisfies the front-door criterion (d, e) exposure-dependent and outcome-dependent selection bias, respectively and (f) selection bias explicitly as collider bias. X nodes represent exposure, Y represents outcomes, Z represents auxiliary data, and S nodes represent selection.

While Pearl was not the first to describe the back-door criterion, he was the first to derive it in a graphical setting. He also identified a second condition for causal inference in the presence of confounders - the "front-door criterion." This criterion states that a causal effect is identifiable if there is a set of nodes $\mathbf{Z}$ such that there is no direct effect of X on Y except that which is mediated by $\mathbf{Z}$. Figure 5.2c shows a simple case in which there are latent variables affecting $X \rightarrow Y$, but $\mathbf{Z}$ lies on the only directed path from X to Y and X blocks all back-door paths from $\mathbf{Z}$ to Y, so $\mathbf{Z}$ satisfies the front-door criterion (Pearl, 1995a).

Selection bias occurs when units are selected into a study with non-uniform probabilities, and the probability of selection is affected by an exposure or an outcome of interest. Figures 5.2d and 5.2e show the simplest cases of exposure-dependent and outcome-dependent selection bias (where the nodes labeled as S represent selection). A causal effect is identifiable in cases of exposure-dependent selection bias since the only path from S to Y is blocked by X, but not outcome-dependent selection bias in which there is no way to block the path between Y and S.

In practice, selection bias commonly presents as volunteer bias, loss to follow-up, or nonresponse bias. These forms of selection bias all share a common structure where selection is a collider or the descendant of a collider (Munafò et al., 2018). Figure 5.2f shows an example of selection bias explicitly as collider bias, where X and Y are common causes of S, and conditioning on S then opens a spurious back-door path between X and Y making the causal effect unidentifiable.

Estimates of association between common ancestors of a collider will exhibit some amount of bias whether or not the two factors are actually independent, but the direction and magnitude of the bias is determined by the true underlying association (Nguyen et al., 2019). As noted by Munafò et al. (2018), this type of selection bias is particularly concerning in a study like the UK Biobank which seeks to identify associations across a wide array of health factors within a study population selected entirely non-randomly. Luckily, there is a vast literature available on methods for ameliorating selection and confounding biases in such studies.

5.2.5 Admissible sets and conditions for recovery from selection bias

Now that we have outlined the problem of selection bias and how to identify it graphically, we will turn our attention to adjustment methods. Recovering from selection and confounding biases in observational studies is usually contingent upon finding an *admissible set* of nodes in the causal graph (Correa et al., 2018). For a given study and corresponding graph, there may be many admissible sets or none at all, in which case recovering from selection and confounding biases is impossible. If an admissible set can be identified, then causation, association, or prevalence can be estimated with the admissible set and an appropriate adjustment procedure.

An *admissible set* is a set of nodes in a causal graph that, when adjusted for, is sufficient for recovering causal effects and associations from data that suffers from selection bias and confounding bias (Correa et al., 2018). The key assumption underlying the theory of admissible sets and conditions for recovery is that the causal graph has been correctly specified, which is a non-trivial task. An incorrectly-specified causal graph may lead researchers to conclude that a causal effect is unidentifiable when it actually might be, or, more concerningly, that an admissible set is sufficient for recovering a causal effect, when in fact it only appears to be given the misspecified SCM.

The rest of this section will describe how to use SCMs to identify admissible sets for causal, associative, and prevalence settings, and then will briefly discuss practical approaches to specifying SCMs and identifying admissible sets.

In causal inference

Bareinboim et al. (2014) and later Correa et al. (2018) use do-calculus and analysis of SCMs to define general conditions for recovering causal effects $\mathbb{P}(\mathbf{y}|do(\mathbf{x}))$ from confounding and selection bias. Consider a graph G with disjoint sets of nodes

$\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ and additional node S, which represents selection into the sample. We are interested in estimating the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ from data that has been sampled from a population, and $\mathbf{Z}$ is a set of auxiliary covariates observed in the sample alongside $\mathbf{X}$ and $\mathbf{Y}$.

Let the set $\mathbf{T} \subseteq \mathbf{V}$ represent the subset of all variables for which unbiased population data is available. For example, say gender is in set $\mathbf{Z}$, then the unbiased population data for gender would be the unbiased estimate of the distribution of gender in our target population. In many cases, not all of the variables in $\mathbf{Z}$ will have corresponding population data available for adjustment, so let $\mathbf{Z}^T \subseteq \mathbf{Z} \cap \mathbf{T}$ represent the subset of variables of $\mathbf{Z}$ for which population data is also available.

Correa et al. (2018) shows that the pair $\{\mathbf{Z}^T, \mathbf{Z}\}$ is *admissible* for recovering $\mathbb{P}(\mathbf{y}|do(\mathbf{x}))$ if and only if for G, is is possible to re-write the formula for $\mathbb{P}(\mathbf{y}|do(\mathbf{x}))$ from Equation 5.4 in terms of quantities that are actually observed, accounting for selection bias. For example, in most studies, we never observe $P(\mathbf{y}, \mathbf{x}, \mathbf{z})$, but instead observe $P(\mathbf{y}, \mathbf{x}, \mathbf{z}|S = 1)$. In order for a causal effect to be *identifiable*, there must be an admissible set that satisfies

$$\mathbb{P}(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{Z}} \mathbb{P}(\mathbf{y}|\mathbf{x}, \mathbf{z}, S = 1)\mathbb{P}(\mathbf{z} \setminus \mathbf{z}^T|\mathbf{z}^T, S = 1)\mathbb{P}(\mathbf{z}^T) \tag{5.5}$$

It is perhaps easier to explain the requirements of an admissible set in terms of graphical models (Correa et al., 2018):

1. $\mathbf{Z}$ and S must block all back-door paths between $\mathbf{X}$ and $\mathbf{Y}$
2. $\mathbf{Z}^T$ must d-separate $\mathbf{Y}$ from S, $\mathbf{Y} \perp\!\!\!\perp S|\mathbf{Z}^T$

Intuitively, statement 1 prevents introducing new confounding bias that might result from conditioning on elements of $\mathbf{Z}$ and statement 2 adjusts for existing confounding bias by blocking non-causal paths. These conditions also allow for the case that some elements of $\mathbf{Z}$ are only used to adjust for confounding bias, and therefore it is not necessary to observe corresponding unbiased population data.

This definition covers the extreme cases of when no external data for $\mathbf{Z}$ is available and when external data is available for all elements in $\mathbf{Z}$. In the first case where $\mathbf{Z}^T = \mathbf{Z} \cap \mathbf{T} = \emptyset$, an effect is recoverable if and only if

$$\mathbb{P}(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{Z}} \mathbb{P}(\mathbf{y}|\mathbf{x}, \mathbf{z}, S = 1)\mathbb{P}(\mathbf{z}|S = 1) \tag{5.6}$$

In the second case, where $\mathbf{Z} \subseteq \mathbf{T}$, an effect is recoverable if and only if

$$\mathbb{P}(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{Z}} \mathbb{P}(\mathbf{y}|\mathbf{x}, \mathbf{z}, S = 1)\mathbb{P}(\mathbf{z}) \tag{5.7}$$

To illustrate this, let us again refer to the example graph G in Figure 5.1. Say we want to identify the causal effect of $X = x$ on Y, which we can write as

$$P(Y|do(x)) = \sum_{z} \sum_{w} P(Y, Z, W|x)$$

and that W is an indicator for selection. Therefore, in reality we observe $P(Y, Z|x, W = 1)$. However, Z and W are d-separated in G by X, as are Y and W, so $Y, Z \perp\!\!\!\perp W|X$. Therefore, using conditional independence, we can rewrite the observed distribution as

$$P(Y, Z|x, W = 1) = P(Y, Z|x)$$

Integrating out W from the causal effect formula above, we are left with

$$P(Y|do(x)) = \sum_{z} P(Y, Z|x)$$

, which is equivalent to the observed distribution. As we are able to use conditional independence rules to re-express the causal effect formula in terms of observed quantities, the causal effect in G is *identifiable*. This is a case in which no external data is required for recovery, since the selection mechanism was independent of the auxiliary set Z.

In association studies

If we are not interested in drawing causal inferences, and instead are only interested in the association between $\mathbf{X}$ and $\mathbf{Y}$, $\mathbb{P}(\mathbf{y}|\mathbf{x})$, then Bareinboim et al. (2014) show we can relax the conditions described in the previous section. Specifically, we do not have to control as stringently for possible confounding. We can apply the same method to developing conditions for recoverability that we used in the causal inference case. Namely, an association is recoverable if we can express it in terms of observed quantities:

$$\mathbb{P}(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{Z}} \mathbb{P}(\mathbf{y}|\mathbf{x}, \mathbf{z}, S = 1)\mathbb{P}(\mathbf{z}|\mathbf{x}) \tag{5.8}$$

$\mathbf{Z}$ is admissible if it blocks paths between $\mathbf{Y}$ and the selection mechanism S such that $(\mathbf{Y} \perp\!\!\!\perp S|\mathbf{Z}, \mathbf{X})$. The same logic and conditional independence rules demonstrated for causal inference in graph G can be applied here to determine whether associative conclusions are identifiable.

Prevalence estimates

An even simpler task is estimating a population quantity from biased data. In that case, we are interested in estimating the unconditional distribution of $\mathbf{Y}$, and can simplify the generalized condition in (5.5) even further to

$$\mathbb{P}(\mathbf{y}) = \sum_{\mathbf{Z}} \mathbb{P}(\mathbf{y}|\mathbf{z}, S = 1)\mathbb{P}(\mathbf{z}) \tag{5.9}$$

$\mathbf{Z}$ is admissible if it blocks paths between $\mathbf{Y}$ and the selection mechanism S such that $(\mathbf{Y} \perp\!\!\!\perp S|\mathbf{Z})$ (Bareinboim et al., 2014).

Identification of admissible sets

A single SCM may have multiple admissible sets, or none at all. In a simple DAG, admissible sets can often be identified visually by the researcher. However, systematic approaches are necessary in more complex settings. Searching a causal graph systematically for admissible sets will take exponential time, so other algorithms have been developed to search more efficiently for such sets (Correa et al., 2018). Correa et al. propose a polynomial decay algorithm, a class of algorithms which will produce the first solution (of potentially many) or fail in polynomial time. Egami and Hartman (2021) takes an alternate, empirically-driven approach that estimates the conditional dependence structure with an undirected random graph and then searches the resulting graph for the minimum set that satisfies the researcher's requirements.

In the non-causal survey setting, Caughey and Hartman (2017) propose selecting adjustment variables using regularized regression, like LASSO. The set of adjustment variables is the union of the sets of variables with non-zero coefficients in two regressions: one of a selection indicator on covariates for which population data is available, and another of an outcome of interest on covariates available in the sample. However, an important caveat to empirical identification methods is that they rely on the assumption that all relevant quantities have been observed, or that there are no unobserved confounders. This assumption is widely made in practice, but is crucial to evaluate critically.

In moving to more empirical approaches, it is important to keep in mind that a valid admissible set is defined not only by the variables that are sufficient for adjustment, but also those that must be excluded from the set to avoid potentially introducing collider bias during adjustment. Empirical methods may not adequately exclude such variables.

Example in the UK Biobank

To illustrate these ideas, let us consider a simple example in the context of the UK Biobank. Figure 5.3 shows an SCM in which Y is an individual's hippocampal volume, X is their socioeconomic status (SES), Z is the subject's level of dementia and S is selection into the study. The paths in the SCM can be described as follows:

- $Y \rightarrow Z$: Decreased hippocampal volume is leads to higher rates of dementia
- $X \rightarrow S$: Individuals of lower SES may face more barriers to participation in the study, like cost of transportation or inability to take time off of work
- $Z \rightarrow S$: Individuals with dementia may be unable to consent to participating in the UK Biobank, or may be less likely than those without dementia to find transportation to a UK Biobank imaging appointment
- $X \leftrightarrow Z$ collider bias induced by conditioning on selection into the study - knowing that someone is of low-SES and was observed in the UK Biobank implies that they are less like than the average adult to exhibit signs of dementia. It is more likely that a participant only had to overcome one hurdle to participation, rather than two.

We observe X, Y and Z under selection bias, and our goal is to estimate the true association between SES and hippocampal volume, $P(Y|X)$. According to the conditions outlined above, in order to recover $P(Y|X)$, we need to be able to write it in terms of the observed quantities under selection bias. First, we re-write the conditional probability to include auxiliary variable Z,

$$P(Y|X) = \sum_z P(Y|X, Z)P(Z)$$

From the DAG, we can see that Z d-separates Y and S, such that $(Y \perp\!\!\!\perp S|Z)$, which means that we can re-write the probability above as

$$P(Y|X) = \sum_z P(Y|X, Z, S = 1)P(Z)$$

The SCM shows that S and Z are clearly dependent, so we cannot do the same for $P(Z)$. Therefore, $P(Y|X)$ is recoverable only if we have external population data for Z, perhaps an estimate of the proportion of UK adults suffering from dementia, and Z is an admissible set. Without this adjustment, our estimate of the association between SES and hippocampal volume will be biased, even if there is no true relationship between the two quantities. More specifically, the collider bias

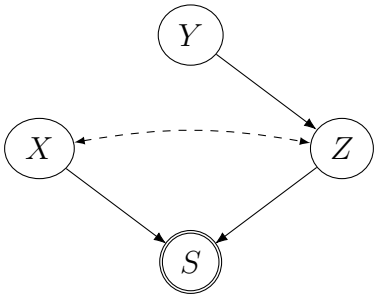


Figure 5.3: Example of selection bias in the UK Biobank. The solid lines represent direct dependence between nodes. The dotted line represents an association induce by collider bias from conditioning on selection into the study, S.

from conditioning on selection implies that someone of low-SES who participates in the UK Biobank is less likely to exhibit signs of dementia, and therefore also less likely to have decreased hippocampal volume. Without adjusting for the true levels of dementia in the population, we might be led to (mistakenly) conclude that lower-SES is associated with larger hippocampal volume, and therefore lower levels of dementia than individuals of high-SES.

5.2.6 Adjustment procedures

Once the researcher has specified the correct causal model and identified a valid admissible set, the task turns to actually estimating the causal effect, associative relationship, or prevalence of interest, $\mathbb{P}(\mathbf{y}|do(\mathbf{x}))$, $\mathbb{P}(\mathbf{y}|\mathbf{x})$, or $\mathbb{P}(\mathbf{y})$, respectively. Equation (5.5) gives an estimator of the causal effect from observed data, however requires estimating $\mathbb{P}(\mathbf{y}|\mathbf{x}, \mathbf{z}, S = 1)$ in each stratum defined by combinations of levels in the auxiliary variables $\mathbf{Z}$. The case described in (5.7) where external data is available for all $\mathbf{Z}$ is exactly **post-stratification**, a common method of weighting non-representative survey responses.

However, as the set $\mathbf{Z}$ increases in size, or if $\mathbf{Z}$ includes continuous variables, this method is infeasible in most practical settings. With many or continuous elements in $\mathbf{Z}$, some strata defined by $\mathbf{Z}$ that exist in the population will not be observed in the sample data, making it impossible to estimate $\mathbb{P}(\mathbf{y}|\mathbf{x}, \mathbf{z}, S = 1)$ in that stratum, and thus the causal effect is unrecoverable. This is the critical problem that the adjustment procedures described in the rest of this section seek to solve.

Inverse probability weighting (IPW)

A broad class of methods that generalizes post-stratification is called **inverse probability weighting** (IPW). These methods rely on a particular re-expression

of (5.5) (Correa et al., 2018):

$$\mathbb{P}(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{Z}} \mathbb{P}(\mathbf{y}|\mathbf{x},\mathbf{z},S=1)\mathbb{P}(\mathbf{z}\setminus\mathbf{z}^T|\mathbf{z}^T,S=1)\mathbb{P}(\mathbf{z}^T) \tag{5.10}$$

$$= \sum_{\mathbf{Z}} \frac{\mathbb{P}(\mathbf{y},\mathbf{x},\mathbf{z}|S=1)}{\mathbb{P}(\mathbf{x}|\mathbf{z},S=1)} \frac{\mathbb{P}(\mathbf{z}^T)}{\mathbb{P}(\mathbf{z}^T|S=1)} \tag{5.11}$$

$$= \sum_{\mathbf{Z}} \frac{\mathbb{P}(\mathbf{y},\mathbf{x},\mathbf{z}|S=1)}{\mathbb{P}(\mathbf{x}|\mathbf{z},S=1)} \frac{\mathbb{P}(S=1)}{\mathbb{P}(S=1|\mathbf{z}^T)} \tag{5.12}$$

Using simple laws of conditional probability, we are able to re-express the estimator for a causal effect in terms of a few interpretable quantities:

- $\mathbb{P}(\mathbf{y},\mathbf{x},\mathbf{z}|S=1)$, the joint distribution of $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{Z}$ under selection bias
- $\mathbb{P}(\mathbf{x}|\mathbf{z},S=1)$, the probability of treatment given covariates in the selection-biased sample data. The conditional distribution $\mathbb{P}(\mathbf{x}|\mathbf{z})$ is commonly known as the *propensity score*
- $\mathbb{P}(S=1)/\mathbb{P}(S=1|\mathbf{z}^T)$, the *inverse probability-of-selection weight*, where $\mathbb{P}(S=1|\mathbf{z}^T)$ denotes the probability of being selected into the sample conditional on unbiased external data. $\mathbb{P}(S=1)$ is the overall probability of being selected into the study which is simply the number of study participants divided by the size of the population and acts as a normalizing constant. Therefore, in practice, we are mainly interested in the denominator $\mathbb{P}(S=1|\mathbf{z}^T)$.

Correa et al. (2018) leverages this re-expression of $\mathbb{P}(\mathbf{y}|do(\mathbf{x}))$ to derive an unbiased estimator of the causal effect in the presence of selection and confounding bias, using the probability of selection, $\mathbb{P}(S=1|\mathbf{z^T})$, and the propensity score $\mathbb{P}(\mathbf{x}|\mathbf{z},S=1)$.

Say we observe $\{\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i\}_{i=1}^n$ under selection bias and some unbiased sample of $\mathbb{P}(\mathbf{Z})$ from a population of size N. Let w_i^C denote the estimated propensity score under selection bias and w_i^S the IPSW for the i^{th} subject:

$$w_i^C = 1/\hat{\mathbb{P}}(\mathbf{X}_i|\mathbf{Z}_i,S=1)$$

$$w_i^S = \hat{\mathbb{P}}(S=1)/\hat{\mathbb{P}}(S=1|\mathbf{Z}_i^T) = (n/N)/\hat{\mathbb{P}}(S=1|\mathbf{Z}_i^T)$$

then the estimator for $\mathbf{Y}$ under treatment condition $\mathbf{X}=\mathbf{x}$ is

$$\hat{\mu} = E[\mathbf{Y}|do(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^{n} w_i^C w_i^S I_{\mathbf{X}_i=\mathbf{x}}\mathbf{Y}_i \tag{5.13}$$

where $I_{\mathbf{X}_i=\mathbf{x}}$ is the indicator for whether the i^{th} subject received treatment condition $\mathbf{x}$.

Intuitively, IPSW can be thought of as creating a pseudo-population which is unaffected by selection bias (Hernán et al., 2004). For example, say that in a population of 100 people, there are 8 women under 30 years old (8% of people). We observe 10 people from the population, one of whom is a woman under 30 (10% of the sample). The overall probability of selection $P(S = 1)$ is $10/100 = 0.1$, and the probability of selection given that an individual was a woman under 30 is $P(S = 1|\mathbf{z}) = 1/8 = 0.125$. Therefore, the woman in the sample is assigned a weight of $w_i^s = 0.1/0.125 = 0.8$. Out of the sample of 10 people, the woman under 30, after weighting, represents 8% of the sample, which is the same proportion as in the target population. If we then rescale the sample weights to have a total weight equal to that of the population by multiplying all the weights by the ratio of the population size to the sample size $N/n = 100/10 = 10$, then the woman under 30 has a weight of 8. This reflects the fact that her responses represent those of all 8 women under 30 in the target population.

Calibration

Within the context of IPW, there are many methods for actually estimating IPW weights $\mathbf{w}^S$. *Calibration* is a general family of methods for producing IPW weights using auxiliary variables. *Post-stratification*, mentioned above, and *raking* are two of the most commonly-used calibration methods.

Consider a population $U = \{1, ..., j, ...N\}$ and a sample drawn from that population. Let $s_j = 1$ if unit j is sampled and 0 otherwise, where $\sum_{j=1}^{N} s_j = n$, and let $S = \{j \in U | s_j = 1\}$. The probability of a unit j being selected for the sample is π_j and is assumed to be strictly positive. We observe some outcome of interest y_j and K auxiliary variables $\mathbf{z}_j = \{z_{j1}, ..., z_{jk}, ..., z_{jK}\}$ for each subject in the sample S. We assume that the population totals of the auxiliary variables, $\mathbf{t_z} = \sum_U \mathbf{z}_j$, are known from some unbiased source, and our goal is to estimate the population total, or prevalence, of y, $t_y = \sum_U y_j$ from the sample data.

The classic estimator of t_y is the Horvitz-Thomson estimator which uses weights based on the *design probability of inclusion* in the sample π_j:

$$\hat{t}_y^d = \sum_{j \in S} y_j/\pi_j = \sum_{j \in S} d_j y_j$$

where d_j is the "design weight" and $d_j = 1/\pi_j$. The design probability of inclusion π_j is equivalent to the denominator of the IPSW, $P(S = 1|\mathbf{Z}_j^T)$, if everyone selected into the study by design is actually and completely observed. However, in practice, for example due to non-response or lack of follow-up, this is almost never the

case. Therefore *calibration* estimators seek to modify the H-T estimator using auxiliary data to account for this reality.

Calibration generates another set of weights for sampled units that are as close as possible to the design weights d_i subject to constraints, or calibration equations, defined by the known population totals of auxiliary variables. These new weights approximate the IPSW w_j^S (times a proportionality constant N/n):

$$w_j^{\text{calibration}} = (N/n)\hat{w}_j^S = 1/\hat{P}(S = 1|\mathbf{z}_j^T)$$

The design weights are not always known, for example in a volunteer-based study like the UK Biobank, in which case equal probability of selection is assumed and design weights are set to N/n. Calibration equations generally have the form

$$\mathbf{t}_z = \sum_{j \in S} w_j \mathbf{z}_j$$

Specific calibration methods are defined by the metric, $G(\mathbf{w}, \mathbf{d})$, used to measure distance between the calibration and design weights, and by the form of the auxiliary data available. For example, the **general regression estimator** (GREG) minimizes the Euclidean distance between $G(w_i, d_i) \propto (w_i - d_i)^2$ while **generalized raking methods** minimize the Kullback-Leibler divergence $G(w_i, d_i) \propto w_i \log(w_i/d_i)$ (Deville and Särndal, 1992). Generalized raking methods, like raking and post-stratification, are also defined by the specific form of the auxiliary information used (i.e. whether the auxiliary information is in the form of marginal or joint distributions) (Deville et al., 1993).

Raking and **post-stratification** are the two most common generalized raking methods. They are highly related, and are in many ways inverses of one another. Post-stratification adjusts the sample joint distribution of auxiliary variables $\mathbf{z}$ to match the population joint distribution. In contrast, raking (or *iterative proportional fitting*), considers each auxiliary variable separately, adjusting the sample marginal distributions of auxiliary variables to match the population marginal distributions. Post-stratification weights can be estimated with a single function evaluation while raking weights are fit iteratively, rotating through each of the specified auxiliary variables until the weights converge according to a tolerance threshold. Post-stratification considers the full joint distribution of auxiliary variables, while raking implicitly assumes independence of auxiliary variables. Post-stratification fails when the strata defined by the joint distribution of the auxiliary variables become too small such that they contain only a few, if any, sample observations. Though a solution set of post-stratification weights is guaranteed, as the probability of

inclusion in a stratum approaches 0, post-stratification weights become extreme or undefined. Raking has no such shortcoming, though there is no guarantee that the algorithm will converge to a set of weights that satisfies all constraints.

Depending on the requirements of a particular application, it is often preferable to use an intermediate procedure, such as one that rakes to a few joint distributions that the researcher believes are important, and the marginal distributions of the remaining variables. This type of procedure leverages the strengths of each raking and post-stratification - it can account for important interactions and also use information from a large number of auxiliary variables.

Calibration has many desirable properties, such as the guarantee that the sample marginal distributions of the auxiliary variables will match those in the population (up to a tolerance threshold) and the lack of assumptions needed about the parametric form of the nonresponse. However calibration methods often fail to capture high-degree interactions in auxiliary variables that may be highly predictive of nonresponse (raking), or are quite limited in the number of auxiliary variables that can be used in weighting (post-stratification). Furthermore, calibration methods call for the researcher to manually define calibration equations, a task that requires extensive knowledge of the application area and will likely differ across surveys, even within the same context. Critiques of weighting methods often cite the highly-manual nature of selection of auxiliary variables and specification of calibration equations (Gelman, 2007).

Parametric modeling of selection probabilities

Calibration frames estimation of selection probabilities as an optimization procedure, however it is also possible to directly model the selection probabilities $\mathbb{P}(S = 1|\mathbf{z}^{\mathbf{T}})$ with some parametric model, like a logistic regression. Logistic regression is simple, will not produce negative weights, and can incorporate a large amount of auxiliary data (including high degrees of interactions). The main drawback of directly modeling the selection probabilities is that there is no guarantee that the weighted population totals of auxiliary variables will match known population totals. This mismatch may introduce bias and reduce interpretability of results. This approach also requires an assumption about the parametric form of selection.

Multiple Regression and Post-Stratification (MRP)

Moving beyond IPW, another set of approaches for adjusting for selection bias involves directly modeling the outcome of interest. These methods can be applied to a wider range of estimators than population totals and ratios, which are the main focus of IPW methods, like coefficients or hyperparameters in complex hierarchical models (Gelman, 2007). However, since these methods, by definition, depend on the outcome of interest, they must be applied to each outcome separately, and it becomes more difficult to analyze relationships between outcomes.

Recall that Equation (5.5) shows that in order to estimate causal effects, we need to be able to estimate $\mathbb{P}(\mathbf{y}|\mathbf{x}, \mathbf{z}, S = 1)$ in each stratum defined by combinations of levels in the auxiliary variable $\mathbf{Z}$. However, as the number of strata defined by $\mathbf{Z}$ increases, the number of responses observed in each stratum decreases, and it quickly becomes impossible to estimate $\hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}, \mathbf{z}, S = 1)$ with the stratum mean, as in post-stratification.

Outcome-based methods use parametric models and a potentially large set of auxiliary variables $\mathbf{Z}$ to estimate $\hat{\mathbb{P}}(\mathbf{y}|\mathbf{x}, \mathbf{z}, S = 1)$ in each stratum. These models pool information across strata so are able to estimate outcomes in strata that were not observed in the sample. Then, these methods use post-stratification to project estimated outcomes on to a target population (Gelman, 2007).

One of the most popular of these methods is **Multi-level Regression and Post-stratification** (MRP), which uses a Bayesian hierarchical model to estimate stratum-level outcomes (Park et al., 2004, 2006). Hierarchical models leverage partial-pooling to generate high-quality estimates for small subgroups, falling back on higher group-level data where stratum-specific information is less available. With high-quality estimates of outcomes in small strata, the main drawback of classical post-stratification can be overcome.

MRP has been used successfully in a wide rage of applications, from YouGov's estimates of election outcomes in the UK and US (Rivers, 2017) to the CDC's estimates of prevalence of key health outcomes at small geographies (Wang et al., 2017). The main drawbacks of MRP are the computational cost of fitting a complex hierarchical model, and the domain knowledge required to correctly specify the model.

Though MRP is a powerful tool, it is not entirely suited to our problem at hand - of estimating a variety of generalizable population quantities based on selection-biased data. MRP is more useful when one is interested in precisely estimating a single population quantity within small subgroups of the population where we may not have many observations. The UK Biobank is unusual in that it

is so large that we do actually observe a sizable number of subjects in subgroups that MRP would usually be used to estimate. Though we do not consider MRP here, it would be interesting to see it applied to another problem involving UK Biobank data in future research.

5.2.7 Variance of weighted estimators

While the methods discussed in Section 5.2.6 decrease the impact of selection bias on estimators, they generally also increase the variance of those estimators. Variance of weighted estimators depends on the values of the weights and also on the procedure used to estimate them (Lu and Gelman, 2003). Even if two adjustment procedures produce numerically identical sets of weights, the variance of the resulting estimators may differ based on the properties of each procedure.

Further complicating variance estimation is that weights are not fixed for individuals given a particular procedure, but depend on the composition of the sample to which that individual belongs. In practice, weighting methods are often layered and adapted to fit the context of a particular application, which makes it difficult to develop a generalizable estimator of variance. Instead, researchers often rely on a few methods of approximation.

Simple Random Sample

In order to motivate these approximations, let us first consider the simple case of a weighted mean calculated from a simple random sample. The weighted mean estimator is given by

$$\hat{\theta} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

In a simple random sample (SRS), we assume that units are selected from the population of interest with equal probability $\pi_i = 1/N$. In an SRS, weights w_i are design-based, so are fixed before a sample is drawn and are equal for all units $w_i \propto 1/n \; \forall \; i \in (1, \ldots, n)$. In this case, we will assume that weights have been re-scaled such that $\sum_{i=1}^{n} w_i = 1$. With these assumptions, we can write the variance of the estimator $\hat{\theta}$ as

$$\text{var}_{\text{SRS}}(\hat{\theta}) = \text{var}\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} w_i^2 \text{var}(y_i) = \left(\sum_{i=1}^{n} w_i^2\right) \text{var}(y_1) = \frac{1}{n}\text{var}(y_1)$$

We estimate $\text{var}(y_1)$ with the weighted sample variance

$$\hat{\text{var}}(y_1) \approx \sum_{i=1}^{n} w_i (y_i - \hat{\theta})^2$$

which gives

$$\hat{\text{var}}_{\text{SRS}}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \hat{\theta})^2 = \frac{1}{n^2}\sum_{i=1}^{n}(y_i - \hat{\theta})^2$$

The SRS estimator fixes weights before sampling, so does not adjust for any observed selection bias. Other estimators that do account for selection bias introduce variance from the weighting procedure, so typically have less bias but higher variance than SRS estimators (Lu and Gelman, 2003).

IPW

Instead of an SRS, we could assume that weights are all IPW weights, so are equal to $1/\pi_i$ where π_i is the actual probability of selection for unit i, and depends on the observed sample. Sampling is most often done without replacement, so the variance of $\hat{\theta}$ depends on the joint inclusion probabilities of each pair of observations, making exact calculation a non-trivial task. Therefore, variance is often estimated as

$$\hat{\text{var}}_{\text{IPW}}(\hat{\theta}) = \sum_{i=1}^{n} w_i^2(y_i - \hat{\theta})^2$$

where we again assume that w_i have been scaled such that $\sum_{i=1}^{n} w_i = 1$ (Lu and Gelman, 2003). In practice, adjustment procedures are often more complex than straightforward IPW, and steps are taken to control the variance of w_i, so this method tends to overestimate the variance of $\hat{\theta}$.

Empirical methods

Due to the complexity of deriving variance formulas even for the simplest weighting procedures, researchers generally rely on more empirical approaches to variance estimation. One such procedure is jackknife variance estimation, where weights are estimated the weighted estimator $\hat{\theta}$ evaluated n times, once excluding each observation in the sample. The variance of $\hat{\theta}$ is estimated by the sample variance across the n iterations. Another procedure used for variance estimation is bootstrapping, in which samples of size n are drawn with replacement from the observed sample, weighted, used to calculate $\hat{\theta}_w$, and the sample variance of $\hat{\theta}_w$ used to estimate $\text{var}(\hat{\theta}_w)$.

Design effect

The design effect of a set of weights measures how much an adjustment procedure increases the variance of a weighted estimator relative to an estimator that assumes an SRS. The design effect for a set of weights is given by the ratio of the variance of weighted estimator $\hat{\theta}_w$ relative to the variance of the SRS estimator $\hat{\theta}_{\text{SRS}}$

$$\text{deff}(\mathbf{w}) = \frac{\text{var}(\hat{\theta}_{\text{w}})}{\text{var}(\hat{\theta}_{\text{SRS}})}$$

Kish (1992) shows that the design effect can also be written in terms of the coefficient of variation $CV(\mathbf{w})$, which is the ratio of the variance of the set of weights $\mathbf{w}$ to the square of the mean:

$$\text{deff}(\mathbf{w}) = 1 + CV(\mathbf{w})^2 = 1 + \frac{\text{var}(\mathbf{w})}{(\text{mean}(\mathbf{w}))^2}$$

The design effect can also be used to calculate the effective sample size (ESS), or the size of a simple random sample that would have variance equal to that of the weighted sample. For a sample of size n, a weighted sample has variance equal to a simple random sample of size $n/\text{deff}(\mathbf{w})$. When all weights are fixed, as in an SRS, the variance of the weights is 0, the design effect is 1 and $ESS = n$. When the variance of the weights is positive, $\text{deff}(\mathbf{w}) > 1$ and $ESS < n$.

5.3 Methods: Adjustment with BART and Raking

Here we propose a simple two-step procedure that alleviates the manual requirements of calibration equation specification while also guaranteeing that final weighted auxiliary variable totals match known population totals. First, we estimate $\mathbb{P}(S = 1|\mathbf{z}^T)$ for each sample observation directly using a Bayesian Additive Regression Tree (BART), $\hat{\mathbf{p}}_1$. We use a BART due to its ability to handle high degrees of interactions and non-linearities in the auxiliary variables. We denote this first set of weights $\hat{\mathbf{w}}_1 = 1/\hat{\mathbf{p}}_1$. Then, we use raking to find a second set of weights $\hat{\mathbf{w}}_2$ such that the weighted sample totals match the population totals of a key subset of variables and the K-L divergence $G(\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2)$ between the sets of weights is minimized.

This will produce a set of weights that is flexible enough to account for complex relationships between auxiliary data and selection, while also remaining interpretable. This method does require some contextual knowledge to identify the variables for which it is most important that weighted sample totals match those of the population for interpretability, but, unlike normal raking, can account for complex interactions between auxiliary variables without pre-specification.

5.4 Methods: Simulation Study

Though the methods outlined in Section 5.2.6 are firmly-established methods for correcting for selection and confounding bias, their relative performance in applications with complex mechanisms of missingness is less clearly understood. This section gives the methodology of a stimulation study using real UK Biobank data to compare the performance of various adjustment procedures.

In the simulation, we select random subsamples of various sizes from the 20,827 UK Biobank imaging subjects, and use adjustment procedures to estimate known quantities of the UK Biobank imaging population from the biased samples. First, we generate a missingness mechanism based on covariate data $\mathbf{Z}$ from the UK Biobank, and use it assign each subject $j = (1, \ldots, N = 20827)$a probability p_j that they are observed. On each iteration of the simulation, we randomly select a sample of a fixed size n_{obs} with probability proportional to p_j. Then, we adjust that sample using each of the methods under consideration. We perform this simulation 7 times, once for each $n_{obs} \in N * (0.01, 0.02, 0.04, 0.05, 0.075, 0.1, 0.25)$. The full algorithm is described in Algorithm 1.

Once we have generated weights for each sample, we calculate weighted estimates of the following quantities:

- **Brain volume**: total brain volume (gray and white matter), normalized for head size, measured in mm^3 by T1 structural MRI. Brain volumes range from 1,151,700mm^3 to 1,793,910mm^3 with a mean of 1,502,37mm^3,
- **Association between brain volume and age**: β_{age} from the weighted linear regression $Y_{\text{brain volume}} = \beta_0 + \beta_{\text{age}} Z_{\text{age}} + \epsilon$

We will calculate the bias of each estimator, as well as the design effect and distribution bias of each set of weights.

Data

The simulations use UK Biobank imaging data. There are 20,827 subjects that have a recorded MRI brain volume, and will serve as the population for these simulations. We observe the following demographic data for each subject:

- **Age** measured at time of imaging appointment
  - continuous: 40 - 79
  - squared: 40^2 - 79^2

Algorithm 1: Simulation 1

Result: Weighted samples

sample missingness coefficients $\boldsymbol{\beta}$;
calculate probability of missingness p_j for all N subjects $p_j = \text{logit}(\mathbf{Z}_j\boldsymbol{\beta})$;
for $\pi_{obs} \in (0.01, 0.02, 0.04, 0.05, 0.075, 0.1, 0.25)$ **do**
  $n_{sim} = \pi_{obs} * N$;
  for $m \in (1, \dots M = 1000)$ **do**
    select sample of n_{sim} subjects;
    $s_j = 1$ if j^{th} subject is selected where $s_j \sim \text{Bern}(p_j|n_{sim})$;
    for *each adjustment procedure* **do**
      weight sample;
      return *weights*
    end
  end
end

  - discrete (7 categories): 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79

- **Sex** (2 categories): Female, Male
- **Ethnicity** (12 categories): White, White Irish, White other, Asian Indian, Asian Pakistani, Asian Bangladeshi, Asian Other, Black Caribbean, Black African, Mixed, Other, do not know/refused
- **Employment** (8 categories): employed, retired, homemaker, disabled, volunteer, student, unemployed, do not know/refused
- **Occupation** (SOC2010, 11 categories): manager, professional, associate professional, administrative, skilled trades, personal service, sales or customer service, industrial, elementary, unemployed, do not know/refused
- **Highest level of education** (6 categories): college or above (including professional degree), A-levels (or equivalent), O-levels or CSEs, vocational or other, none, do not know/refused
- **Household income** (£1000s, 6 categories): Under £18, £18-£31, £31-£52, £52-£100, Over £100, do not know/refused
- **Household size** (6 categories): 1, 2, 3, 4, 5 or more, do not know/refused
- **Home ownership** (6 categories): Own outright, own with a mortgage, rent from LA, rent from a private landlord, rent free, do not know/refused

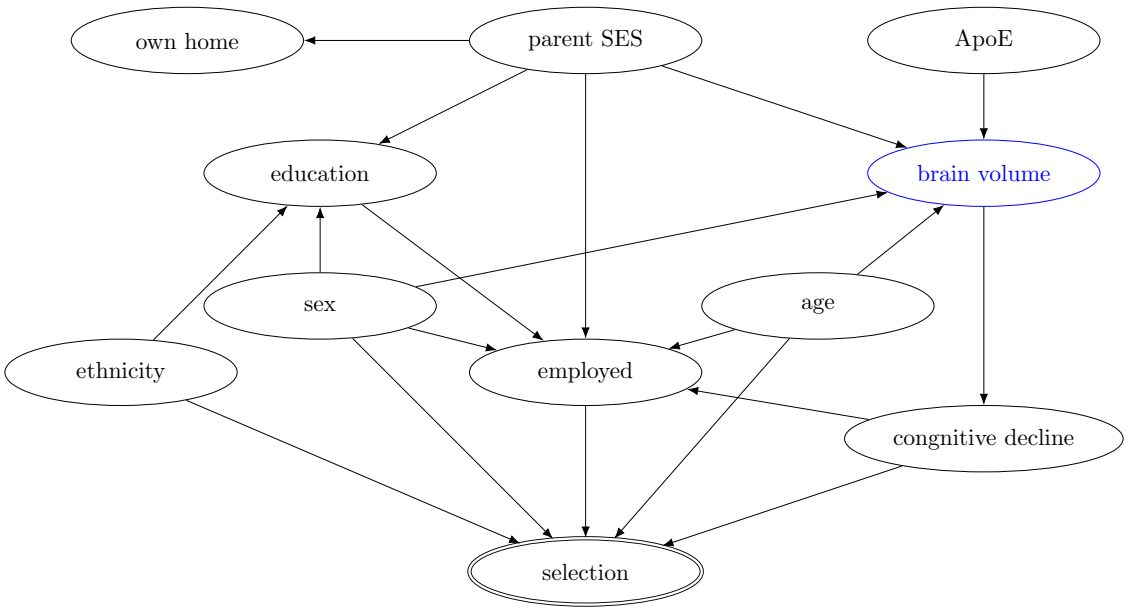
- **Home type** (3 categories): house, flat/apartment/temporary accommodation, do not know/refused

Further explanation of these demographic and outcome variables can be found in Section 5.5.

Simplified SCM of the UK Biobank

Figure 5.4 shows a simplified SCM of key variables observed in the UK Biobank imaging cohort. This SCM is based on assumptions of likely mechanisms that affect selection bias and current understanding of relationships between

Figure 5.4: Simple hypothetical SCM of the UK Biobank imaging cohort data.



Probability of missingness

The simulation requires generation of a missingness mechanism, or a linear combination of demographic variables $\mathbf{Z}$ and randomly-drawn coefficients $\boldsymbol{\beta}$ that is used to estimate a probability of selection p for each subject in the population, $p_j = \text{logit}(\boldsymbol{\beta}\mathbf{Z}_j)$.

We generate a single set of coefficients $\boldsymbol{\beta}$, which we would like to be 1) perfectly recoverable, 2) create significant bias in estimates of outcomes and 3) be realistically complex.

In order to ensure the first criterion, probability of missingness is a function only of variables $\mathbf{Z}$ that can be considered by the weighting procedures (listed in the previous section). For the second criterion, we always include age in the

missingness function, which is well-known to be correlated with brain volume. In the UK Biobank imaging data, age and brain volume have a correlation of -0.55.

The third criterion stems from the result shown in Equation 5.5. We know that if we observe all variables $\mathbf{Z}$ that d-separate $\mathbf{Y}$ and S, and can estimate $\hat{\mathbf{y}}$ in each cell defined by the joint distribution of $\mathbf{Z}$, we will always be able to un-biasedly recover the quantity of interest using post-stratification, and post-stratification will dominate other methods (Caughey and Hartman, 2017). However in most practical applications, $\mathbf{Z}$ is large or includes a continuous variable, making it impossible to estimate $\hat{\mathbf{y}}$ in all cells formed by the joint distribution of $\mathbf{Z}$.

We seek to evaluate adjustment methods under realistic conditions, so would like $\mathbf{Z}$ to be large and contain continuous variables. Age will always be included in order to induce bias (criterion 2). We also ensure that $\mathbf{Z}$ is complex by considering each level of each discrete variables separately by coding all categorical variables as a set of indicator variables, without removing a default reference value. For example, home type is coded as 3 variables, each corresponding to one of the levels (example shown in Table 5.1).

Home type	**house**	**flat_or_apartment**	**refused**
House	1	0	0
Flat or apartment	0	1	0
Do not know/refused	0	0	1

Table 5.1: Example of variable coding for simulating the probability of missingness.

This clearly creates collinearity between predictors, and we will rely on the sparsity of the simulated $\boldsymbol{\beta}$ to ensure that not all levels of a single categorical variables are included in a single function. Furthermore, even if by chance all levels of a variable were included, since we are simply simulating a function for p and not estimating it in a regression, one of the levels could be considered an intercept term.

Real missingness mechanisms are often non-linear. We introduce non-linearity here by including age^2 in $\mathbf{Z}$ and by considering all two-way interactions of demographic variables. We have $(\text{age}, \text{age}^2, 7 \times \text{age categories}, 2 \times \text{sex}, 12 \times \text{ethnicity}, 8 \times \text{employment}, 11 \times \text{occupation}, 6 \times \text{education}, 6 \times \text{income}, 6 \times \text{household size}, 6 \times \text{home ownership}, 3 \times \text{home type}) = 69$ first-order predictors. Considering two-way interactions introduces an additional $\binom{69}{2} = 2346$ predictors, for a total of $K = 2415$ possible predictors. While this may seem extensive, we have only considered a small fraction of the variables available in the UK Biobank, and only one type of non-linearity.

We use a spike-and-slab distribution to simulate the missingness coefficient β_k for all $k \in (1, \ldots, K)$ predictors. The spike and slab distribution is a mixture model, using a Bernoulli distribution to model the probability that random variable is non-zero (the "spike") and a Normal distribution to model the value of the variable given that it is different from 0 (the "slab"). The parameters of the distribution are λ_k, the probability that β_k is non-zero, and μ_k and σ_k^2, the mean and variance of the Normal distribution that β_k follows if non-zero. The hyperparameter α_k indicates if β_k is non-zero.

$$\lambda_k = \begin{cases} 1 \text{ for age} \\ 0.75 \text{ for age}^2 \\ 0.5 \text{ for continuous variables} \\ 0.25 \text{ for first-order binary indicator variables} \\ 0.003 \text{ for interaction variables} \end{cases}$$

$$\alpha_k \sim \text{Bern}(\lambda_k)$$

$$\beta_k \sim \alpha_k \mathcal{N}(\mu_k, \sigma_k^2)$$

The sampled values of $\boldsymbol{\beta}$ used in the simulation study are given in Table 5.6 in the Appendix.

As described in Algorithm 1, once the $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ have been sampled, we calculate p_j for all $j \in (1, \ldots, N)$ as $p_j = \text{logit}(\mathbf{Z}_j \boldsymbol{\beta})$. Then, holding the sample size $n_{sim} = N\pi_{obs}$ fixed, we draw a sample from the population. Let s_j be the indicator for the j^{th} person being sampled, then $s_j \sim \text{Bern}(p_j | n_{sim})$. We consider a range of proportions observed π_{obs} from 0.01 to 0.25 in order to evaluate how the performance of adjustment procedures varies with the amount of data available.

Adjustment methods

Each simulation will compare 6 adjustment methods: 1) post-stratification 2) raking 3) calibration 4) raking with LASSO variable selection 5) logistic regression to estimate p_i and 6) BART with raking. The theory behind methods is given in Section 5.2.6, this section will discuss the practical implementation of methods in the simulation.

Post-stratification

One of the main disadvantages of post-stratification is the need for the researcher to select variables with which to strata for weighting that capture as much of the missingness mechanism as possible without resulting in tiny cells that contain few or no observations. This is typically a manual process that relies on domain

knowledge, however, that is impractical in this setting. We use a random forest to simulate the domain knowledge typically used for variable selection. Specifically, we use a random forest (from the `randomForest` package in R) to predict the vector of selection indicators `s` using the discrete Z_k in their categorical (not binary indicator) forms. We exclude the continuous Z_k as post-stratification can only use discrete variables, and implicitly considers variable interactions, so there is no need to explicitly specify them in the random forest.

The random forest generates a measure of variable importance. Beginning with the variable deemed most important, we add on additional variables of decreasing importance until the joint distribution of the selected variables creates cells that have no observations in the sample and make up, in aggregate, no more than 1% of the population. For example, consider if the two most important variables were age buckets and gender and no cells based on the interaction of the two variables were empty in the sample. We would add the next important variable, say income, and re-calculate the number of observations in each cell of the joint distribution of all 3 variables. Say that there are 2 cells for which we lack sample observations, but combined, they make up only 0.5% of the population. This is acceptable, and we move on to consider the next most-important variable, employment status. The joint distribution of all 4 variables contains 10 empty sample cells, which make up 1.5% of the population, which is over out threshold. Therefore, we eliminate the 4^{th} variable and proceed to post-stratification with the first three.

Calibration

Calibration minimizes the distance between prior weights, set to 1 here, and new weights that satisfy a set of constraints. Constraints are defined as functions of auxiliary variables, often population totals. Here we consider 3 sets of constraints:

- Population size in the form of $N = \sum_{i=1}^{n} w_i$
- Sums of continuous variables (age and age^2) in the form of $\sum_{j=1}^{N} z_j = \sum_{i=1}^{n} w_i z_i$
- Counts across levels of the 10 categorical variables. For this constraint, we transform categorical variable Z_k that has L_k levels into $L_k - 1$ binary indicator variables, dropping one level per variable to avoid collinearity. Constraints take the form

$$\sum_{j=1}^{N} I\{z_j = l_k\} = \sum_{i=1}^{n} w_i I\{z_i = l_k\}$$

for all $l \in (1, \ldots, L_k - 1)$ levels of Z_k, and all Z_k in the $k \in (1, \ldots, 10)$ categorical variables.

In total, there are $(1 \times \text{population total} + 2 \times \text{age totals} + 6 \times \text{age categories} + 1 \times \text{sex} + 11 \times \text{ethnicity}, 7 \times \text{employment} + 10 \times \text{occupation} + 5 \times \text{education} + 5 \times \text{income} + 5 \times \text{household size} + 5 \times \text{home ownership} + 2 \times \text{home type}) = 60$ constraints to consider in each calibration estimation.

With so many constraints, calibration can perform poorly - either by producing extreme weights, or by failing to converge altogether. In order to prevent this, we eliminate constraints that apply to levels of discrete variables that make up less than 2% of the sample or the population. By eliminating these constraints, we effectively pool these small levels with the reference level for that categorical variable, which has been chosen somewhat arbitrarily. Generally in a real application setting, the researcher would manually pool small strata based on domain knowledge.

The second step we take to aid convergence is calibrating in two stages. We split the constraints into two groups based on the number of observations in the population that are represented by that level, and calibrate the sample first with the constraints for smaller population subgroups, then use the resulting weight as the prior weight for a second round of calibration with the larger constraints. We calibrate the smaller groups first so that the final weighted sample exhibits less overall error in marginal distributions of auxiliary variables. This approach is ad-hoc, and should be tested against other methods for dealing with a large number of constraints in future research.

The last step we take to aid convergence in calibration is the specification of algorithm parameters - namely the tolerance threshold and the maximum number of iterations that the algorithm will be allowed. The tolerance threshold ϵ is the threshold that determines when the weighted sample total matches the population total:

$$\left| \sum_{j=1}^{N} I\{z_j = l_k\} - \sum_{i=1}^{n} w_i I\{z_i = l_k\} \right| < \epsilon$$

By default, ϵ is set to $1e-7*N$, however we raise this threshold to $\epsilon = 0.0003*N$. This is primarily based on anecdotal observations of cases when calibration failed to converge despite other precautions taken. Further research should explore performance of calibration weights as a function of this tolerance. Lastly, the maximum number of iterations was increased from the default of 50 to 5000.

We used the `calibrate` function from the `survey` package in R to fit calibration weights.

Raking

As discussed in Section 5.2.6, raking is a specific form of calibration in which the constraints are marginal distributions of categorical variables. As in calibration,

raking will fail to converge when there are a large number of constraints or when constraints include discrete variables with levels representing small population subgroups. In order to avoid this, we eliminated members of the population that were absent in the sample, as is standard practice in post-stratification. For example, if the sample contained no observations with ages between 40 and 45, we eliminated all members of the population between those ages, and calculated target population marginal distributions based on the remaining members of the population. This clearly presents a problem if we are forced to drop small, but potentially important, population subgroups to fit rake weights. In most real applications, the researcher would pool levels of the population that were missing from the sample with other levels based on domain knowledge. However this is a manual process, and a large drawback of raking, which we would like this simulation to capture.

We used the `rake` algorithm in the `survey` package in R to fit rake weights. This implementation allows for the specification of a tolerance threshold ϵ and of a maximum number of iterations. As in calibration, we set the tolerance threshold to be 0.0003, and the maximum number of iterations to be 5000.

Raking with LASSO variable selection

This method approaches adjustment as a variable selection problem. Raking can consider a large set of auxiliary variables, while post-stratification quickly becomes impossible when more than a few auxiliary variables are considered. On the other hand, post-stratification can account for interactions in the missingness mechanism, while raking only adjusts the marginal distributions of the auxiliary variables. This method attempts to leverage the advantages of each raking and post-stratification by adjusting on the minimum set of interactions significantly related to the outcome, in this case total brain volume, or probability of selection. LASSOs are used to select significant predictors of brain volume or selection, and then the sample is raked only to the marginal distributions of those variables. We excluded age and age^2 from consideration, as we cannot rake to the population margins of continuous variables. This left $K = (69 - 2) + \binom{69-2}{2} = 2278$ possible predictors.

For the LASSO predicting selection, we use all 20,827 observations in the population and predict the binary outcome s_j which is 1 if the j^{th} observation was selected in the sample, and 0 otherwise. For the outcome LASSO, we restrict the training data to the sample and assume that the outcome, brain volume, is normally distributed conditional on covariates $\mathbf{X}$. We use `cv.glmnet` from the `glmnet` package in R, and 5-fold cross-validation.

In practice, there are a few challenges to address. First, any variable selected by one of the LASSOs will be used in raking, so there must be enough observations of

each level in both the sample and population. We set that threshold to be 1%, thus eliminating any variables that create population or sample cells below this threshold.

Second, in order to decrease the computation required to fit each LASSO, we reduce the number of λ penalty values considered. By default, `cv.glmnet` considers 100 values of λ from $\lambda_{\min}$ to $\lambda_{\max}$, equally-spaced on the log scale. $\lambda_{\max}$ is set such that when $\lambda > \lambda_{\max}$, all coefficients are 0. Friedman et al. (2010) show that β_k will be 0 if $\frac{1}{N}|\langle x_k, y\rangle| < \lambda\alpha$, and therefore all coefficients will be 0 when $N\lambda_{\max} = \max_k |\langle x_k, y\rangle|$. $\lambda_{\min}$ is set to $0.001\lambda_{\max}$. We take a similar approach but consider only 20 values between $\lambda_{\min}$ and $\lambda_{\max}$ instead of 100. The final set of variables was selected using the largest λ such that the cross-validated error lies within 1 standard deviation of the minimum cross-validated error. This criterion is often used to prevent overfitting (Friedman et al., 2010).

Last, the LASSOs frequently identified 30 or more significant variables, which, when considered all at once, caused the raking algorithm to fail to converge. To prevent this, we capped the number of raking variables at 50 and then raked using sets of 20 variables at a time, from least important to most important. On each raking iteration, the weights from the previous iteration were used as the prior weights. Variable importance was determined first by whether the variable appeared in one or both LASSOs, and second by the relative importance of the variable within the LASSO. Due to the difference in the scale of the outcomes across the two LASSOs, we relied on relative coefficient size rather than absolute coefficient size.

Once raking variables and groups were identified, we implemented raking using the same settings as the straight raking approach described above.

Logistic regression

Logistic regression for selection bias adjustment takes a different approach than the previous methods by attempting to directly estimate the probability of selection instead of attempting to make the sample margins match the population margins. In fact, this method makes no attempt to match population margins.

We use a logistic regression LASSO for variable selection. Similarly to the LASSOs used in the previous method, we used `cv.glmnet` with 5-fold cross-validation and custom values of λ. The optimal λ was selected using the 1 standard deviation criterion, and the model was then re-fit using only significant predictors and no penalty parameter to avoid coefficient shrinkage in the final predictions. Occasionally, the LASSO would fail to select any significant variables. In that case, a simple logistic regression was fit using all 69 first-order predictors and no interactions.

The weights from the resulting logistic regression are $1/\hat{p}_j$ where $\hat{p}_j$ is the modeled probability of selection.

BART and raking

BART and raking is similar to the previous method in that it attempts to directly estimate the probability of selection. The additional raking step then ensures that sample marginal distributions of key variables match those of the population.

We use the `BayesTree` package in R to fit the BART on all 69 first-order terms, including age and age^2. BARTs naturally consider non-linearities and interactions between predictors, so there is no need to manually pre-specify them. We fit a categorical BART with 25 trees, and the estimate the probability of selection for each subject with the mean of 1000 samples from the posterior of the model.

The BART-estimated probabilities serve as our prior probabilities in raking. To avoid over-raking, and potentially increasing the variance unnecessarily, we select a subset of variables for raking. The `BayesTree` implementation of BART does not have a variable importance metric, so we fit a simple random forest from the `randomForest` package to the same data using the 8 categorical covariates instead of the 67 binary covariates, and select the top 5 most important variables from that model for raking. Raking settings were the same as those used for simple raking, described above.

Other considerations

For all methods, weights were re-scaled to have mean 1 to simplify comparison across methods. Occasionally methods would fail to converge despite precautions taken. In that case, a weight of 1 was assigned to all sampled units.

Analysis

We will evaluate adjustment methods based on their ability to eliminate selection bias in average total brain volume relative to the amount of additional variance introduced by weighting. For each sample and method, we will calculate the bias in adjusted total brain volume $bias(t_{bv,w})$, the design effect of the weights deff($\mathbf{w}$) and the distribution bias of the weighted sample marginal distributions DB($\mathbf{w}$). For each method and proportion-sampled combination, we will also calculate the mean squared error (MSE) of adjusted total brain volume:

$$MSE(t_{bv}) = \frac{1}{m}\sum_{m=1}^{M} bias(\mathbf{w}_m)^2$$

The design effect and the MSE measure different facets of the variability of adjustment procedures. The deff depends only on the set of weights, while the MSE is specific to the outcome of interest.

We will evaluate these metrics not only at the sample-wide level, but also within subgroups.

5.4.1 Methods for application to the UK Biobank

With an understanding of the relative benefits of various adjustment procedures from the previous section, we will then apply them to the actual UK Biobank imaging data. We will attempt to estimate the following population characteristics:

- **Prevalence** of smoking
- **Prevalence** of obesity
- **Prevalence** of ApoE e4/e4 phenotype
- **Population mean** total brain volume
- **Association** between age and total brain volume

This list of outcomes represent two broad categories of outcomes with different levels of strictness of conditions for recovery that are both of interest in studies like the UK Biobank: prevalence $\mathbb{P}(\mathbf{Y})$ and association $\mathbb{P}(\mathbf{Y}|\mathbf{X})$. Second, these outcomes have been widely studied, so the literature provides a strong benchmark by which to evaluate our adjustment procedures (Brandon et al., 2018; Fotenos et al., 2008).

To estimate these outcomes, we will apply each of the weighting methods under consideration to the UK Biobank imaging cohort using the 2016 HSE as our target population to define population totals of auxiliary variables (NatCen Social Research, 2018). Then, we will calculate weighted estimators of each quantity. As in the simulation study, the association between total brain volume and age will be calculated using weighted linear regressions.

We will calculate the design effect of each set of weights, and compare estimates to population quantities from the 2016 HSE.

5.5 Methods: Data

This analysis sources data from the UK Biobank and the 2016 HSE. UK Biobank data is publicly accessible upon approval of an application through the UK Biobank Access Management System. HSE data was accessed via the UK Data Service.

The UK Biobank is the selection-biased data that we would like to adjust, and the 2016 HSE is our target population. As we aim to directly compare demographic and health metrics from the UK Biobank to those from the HSE, we recode relevant variables to match across sources as closely as possible. For example, this may include collapsing levels of a certain variable in the UK Biobank because the corresponding variable was collected at a higher level of aggregation in the HSE. We will describe a few such cases in the rest of the section, but full recodes can be found in the GitHub repository https://github.com/vcbradley/ukb-selection-bias.

5.5.1 Methods for comparing populations

In order to evaluate the extent of the selection bias in the UK Biobank imaging cohort, we compare key demographic and health indicators to those from high-quality studies of the target population. The specific data used is discussed in more detail in Section 5.5. Our main metric of comparison is the proportion of each population that falls into levels of relevant categorical variables. For variables for which we have continuous data, like age, we create relevant buckets (i.e. 40 to 45, 45-50, etc.) and compare distributions across those buckets. Wherever possible, we also compare means and variances of continuous variables. However, we are highly restricted by the comparable data available for each population due to reporting styles and privacy concerns.

5.5.2 UK Biobank data

The UK Biobank is a national prospective health study of UK adults ages 40-69 at time of recruitment, which occurred between 2006 and 2010. All participants completed extensive questionnaires, underwent physical and mental health examinations, gave biological samples and consented to have their National Health Service (NHS) records accessed by the study. The goal of the study is to collect data on diseases of aging, before onset (Fry et al., 2017).

Additionally, up to 100,000 participants will undergo imaging assessments, including MRI of the brain, heart and abdomen, and full-body bone and joint X-ray. The first subjects were imaged in 2016, and imaging is expected to continue through 2022 (Miller et al., 2016). To date, 34,890 subjects have undergone brain MRI imaging. We restrict our sample to the 20,827 observations that contain complete measurements of T1 total brain volume (grey and white matter), normalized for head size.

Demographic variables, like household income and highest education level, were collected once at baseline, and again at the time of the imaging assessment. As we are interested in correcting potential selection bias in brain MRI data with auxiliary demographic variables, we take the value collected at the time of imaging wherever available. If missing, we take the next most recent observation.

Some variables, like income and home ownership, contain observations where the respondent refused to respond or was unsure of the correct response. We code those cases as "Do not know/Refused" and do not impute response values, and do not exclude them from the data, as doing so may introduce additional bias. The HSE also

contains "Do not know/Refused" values for these variables, so there is a subset for comparison, though the pattern of refusals may be different across the two studies.

Some variables, like age and ethnicity, are coded in multiple ways - once at the most granular level for use in simulation studies, and once at a higher level of aggregation to match the way the information is collected in the HSE. Age, for example, is collected in the UK Biobank as a combination of month and birth year. We impute continuous age as the "age in years" from the 15^{th} of the observed birth month and the date of the imaging appointment. In the simulation studies, we use a continuous version of age, however, the HSE only reports age in 5-year increments. Therefore, we create a discrete version of UK Biobank age to match the HSE variable which is used to weight the Biobank data. In the UK Biobank, subethnicity, like "White Irish" and "Black African" is collected in addition to the larger ethnicity categories: white, black, Asian, mixed, other, and no response. However only the major categories were collected by the HSE. Therefore, we code ethnicity two ways: one at the most granular level to use in simulation studies and one at the higher level of aggregation for weighting to the HSE.

We also consider a small selection of health outcomes:

- **Smoking status** (4 categories): current, previous, never, do not know/refused
- **BMI category** (5 categories): underweight (< 18.5), healthy ($18.5 - 24.9$), overweight ($25 - 29.9$), obese (> 29.9), do not know/refused
- **Ever diagnosed with diabetes by a doctor** (3 categories): yes, no, do not know/refused
- **Ever diagnosed with high blood pressure** (3 categories): yes, no, do not know/refused

These variables were selected because the HSE collects similar outcomes, giving high-quality population prevalence estimates for comparison. These health outcomes were not used in the simulation studies, or to weight the UK Biobank data, but serve as benchmarks for how much healthy volunteer selection bias exists in the UK Biobank and how well weighting methods are able to adjust for it.

5.5.3 Health Survey for England (HSE)

The HSE is an annual survey conducted by the Joint Health Surveys Unit of NatCen Social Research and the Department of Epidemiology and Public Health at University College London (NatCen Social Research, 2018). We use the 2016 data as it was the latest available at the time of this analysis. The UK Biobank imaging study began in 2016, so there is a slight mismatch in time of collection between the UK Biobank data and our target population, and weighted estimates will correspond to a nationally representative 2016 adult population.

The 2016 HSE interviewed 8,011 adults aged 16 and over, and 2,056 children under the age of 16. We restrict our data to the 4,318 adults aged 44-79 as the UK Biobank imaging subjects only fall within that age range. The health metrics that we are interested in comparing to the UK Biobank are only available for a subset of the overall sample, so we further restrict the sample to 2,348 individuals who are aged 44-79 and underwent a nurse interview. We use the HSE-supplied survey weights for the nurse interview subset for all population calculations.

The UK Data Service releases anonymized individual-level results for the HSE, which we use here.

5.6 Results

This section discusses 3 sets of results. First, we compare the UK Biobank imaging cohort to the 2016 HSE to demonstrate the existence of selection bias in the UK Biobank. Then, we give the results of simulation studies designed to test 6 methods for adjusting for selection bias. Last, we apply the 6 adjustment procedures to the real UK Biobank imaging cohort and compare adjusted and unadjusted estimates of key outcomes.

5.6.1 Bias in the Neuro Imaging Cohort

Table 5.2 shows compares the UK Biobank imaging cohort to the 2016 HSE nurse interview sample. The table gives population counts (weighted in the case of the HSE) and the distribution of each study across levels of demographic variables.

Sociodemographic factors The UK Biobank imaging cohort is older than the HSE (44.4% and 28.9% aged 65 or older, respectively), more educated (51.5% v. 27.6%), more white (97.1% v. 89.5%) and more likely to be retired (56% v. 39.4%) or to own a home (73.4% v. 39.5%). Notably, the cohort does not have a gender bias (52% women compared to 51.5% in the HSE), despite the higher participation rates among women found by Fry et al. (2017).

Health characteristics The UK Biobank imaging cohort is healthier than the general population. Only 3.8% are current smokers, compared to 16.2% of the general population. The cohort is less likely to be obese (18.7%) than the HSE (28.5%), to have ever been diagnosed with high blood pressure (22.5% v. 30.8%) or to have ever been diagnosed with diabetes (5.1% v. 9.8%).

Table 5.2: Selection bias in the UK Biobank (UKB) imaging cohort relative to the 2016 HSE nurse interview subsample.

	Count		% of sample		
Level	HSE	UKB	HSE	UKB	% UKB-% HSE
01-Sex					
Female	1436	11243	51.1	52.5	1.4
Male	1373	10164	48.9	47.5	-1.4
02-Age Bucket					
45 to 49	414	834	14.7	3.9	-10.8
50 to 54	449	3016	16.0	14.1	-1.9
55 to 59	352	3600	12.5	16.8	4.3
60 to 64	351	4463	12.5	20.8	8.4
65 to 69	346	5195	12.3	24.3	11.9
70 to 74	250	3424	8.9	16	7.1
75 to 79	217	875	7.7	4.1	-3.6
03-Highest Education					
01-College plus/profesh	776	11018	27.6	51.5	23.8
02-A Levels	693	2617	24.7	12.2	-12.4
03-O Levels/CSEs	676	4854	24.1	22.7	-1.4
04-Vocational/Other	35	1312	1.2	6.1	4.9
05-None	626	1387	22.3	6.5	-15.8
99-DNK/Refused	4	219	0.1	1	0.9
04-Disabled					
01-Yes	152	183	5.4	0.9	-4.6
02-No	2657	21224	94.6	99.1	4.6
05-Employed					
01-Yes	1579	8864	56.2	41.4	-14.8
02-No	1230	12543	43.8	58.6	14.8
06-Homemaker					
01-Yes	170	762	6.1	3.6	-2.5
02-No	2639	20645	93.9	96.4	2.5
07-Retired					
01-Yes	825	11985	29.4	56	26.6
02-No	1984	9422	70.6	44	-26.6
08-Student					
01-Yes	20	79	0.7	0.4	-0.3
02-No	2789	21328	99.3	99.6	0.3
09-Unemployed					
01-Yes	39	126	1.4	0.6	-0.8
02-No	2770	21281	98.6	99.4	0.8
10-Volunteer					
02-No	2809	20338	100.0	95	-5

11-Ethnicity					
01-White	2515	20782	89.5	97.1	7.5
02-Mixed/Other	51	253	1.8	1.2	-0.6
03-Asian	163	215	5.8	1	-4.8
04-Black	80	119	2.8	0.6	-2.3
99-DNK/Refused	1	38	0.0	0.2	0.1
12-Own/Rent House					
01-Own outright	1109	15711	39.5	73.4	33.9
02-Own with mortgage	904	4477	32.2	20.9	-11.3
03-Rent from LA	463	373	16.5	1.7	-14.7
04-Rent private	286	442	10.2	2.1	-8.1
05-Shared	19	44	0.7	0.2	-0.5
06-Rent free	27	88	0.9	0.4	-0.5
99-DNK/Refused	1	272	0.0	1.3	1.2
13-Income					
01-Under 18k	412	2526	14.7	11.8	-2.9
02-18k to 31k	707	5454	25.2	25.5	0.3
03-31k to 52k	471	5872	16.8	27.4	10.7
04-52k to 100k	578	4254	20.6	19.9	-0.7
05-Over 100k	83	1177	3.0	5.5	2.5
06-DNK/Refused	558	-	19.8	-	-
14-Occupation					
01-manager	159	1450	5.7	6.8	1.1
02-professional	327	2154	11.7	10.1	-1.6
03-assoc professional	196	1667	7.0	7.8	0.8
04-admin	179	1195	6.4	5.6	-0.8
05-skilled trades	172	466	6.1	2.2	-3.9
06-personal service	158	470	5.6	2.2	-3.4
07-sales customer service	85	244	3.0	1.1	-1.9
08-industrial	133	262	4.7	1.2	-3.5
09-elementary	169	265	6.0	1.2	-4.8
10-unemployed	1230	12543	43.8	58.6	14.8
99-DNK/Refused	1	691	0.0	3.2	3.2
15-Smoking status					
01-Current	454	824	16.2	3.8	-12.3
02-Previous	994	7182	35.4	33.5	-1.9
03-Never	1360	13186	48.4	61.6	13.2
99-DNK/Refused	0	215	0.0	1	1
16-BMI Bucket					
01-Underweight	15	141	0.5	0.7	0.1
02-Healthy	686	7942	24.4	37.1	12.7
03-Overweight	1045	8813	37.2	41.2	4
04-Obese	800	4013	28.5	18.7	-9.7
99-DNK/Refused	263	498	9.4	2.3	-7
17-Ever diagnosed high BP					
01-Yes	864	4810	30.8	22.5	-8.3
02-No	1941	16096	69.1	75.2	6.1
03-DNK/Refused	5	501	0.2	2.3	2.2
18-Ever diagnosed diabetes					
01-Yes	275	1101	9.8	5.1	-4.7
02-No	2532	20306	90.1	94.9	4.7
99-DNK/Refused	1	-	0.1	-	-

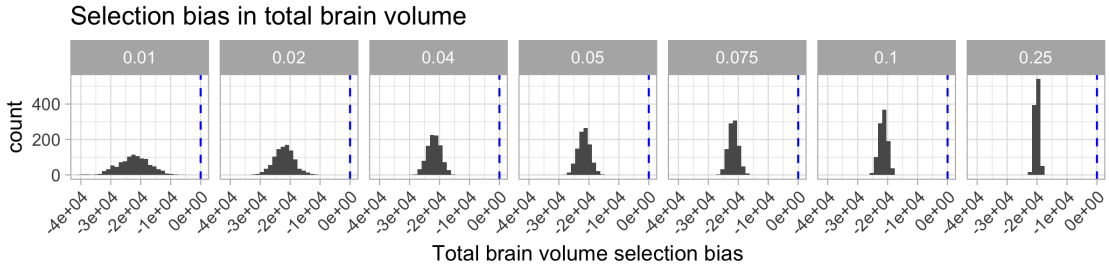


Figure 5.5: Histogram of mean total brain volume in simulated samples. The blue dotted line represents the true population mean total brain volume.

5.6.2 Simulation results

Results from simulation studies are arranged by outcome of interest: total brain volume, association between total brain volume and ApoE, and population composition.

Total brain volume

Figure 5.5 shows the distribution of mean total brain volume in our random subsamples, All samples have mean total brain volumes far below the true population value, showing that we successfully induced selection bias.

Bias Figure 5.6 shows the bias in estimated total brain volume remaining after adjustment across weighting methods and sample sizes. At the smallest sample size (208), the BART method produces the best estimate of population total brain volume, followed by stratification and raking. BART produces unbiased estimates at larger sample sizes, until a slight dip when the sample is a full 25% of the population. At larger sample sizes, 833 and above, post-stratification performs at least as well as BART. The magnitude of the bias in the logit-weighted estimator decreases consistently as sample size increases. Calibration, lasso and raking estimators are not unbiased at any sample size.

Design effect Figure 5.7 shows the design effect of each method across sample sizes. The calibration and logit weights have considerably larger design effects, and design effects that are themselves highly variable. The LASSO method has the smallest design effect across all sample sizes. BART and raking have similarly small design effects, consistent across sample sizes, while post-stratification has small design effects that increase slightly with sample size.

MSE Figure 5.8a shows the log MSE of total brain volume estimators as a function of the proportion of the population sampled. The BART estimator has the lowest MSE for 1% and 2% of the population sampled, but is surpassed by post-stratification at larger sample sizes. The logit and raking estimators perform

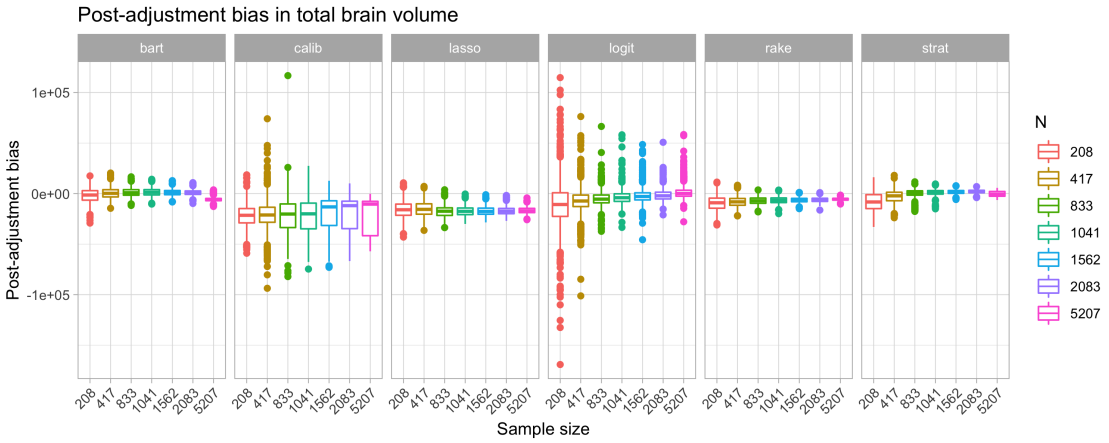


Figure 5.6: Error by method and sample size in simulation studies

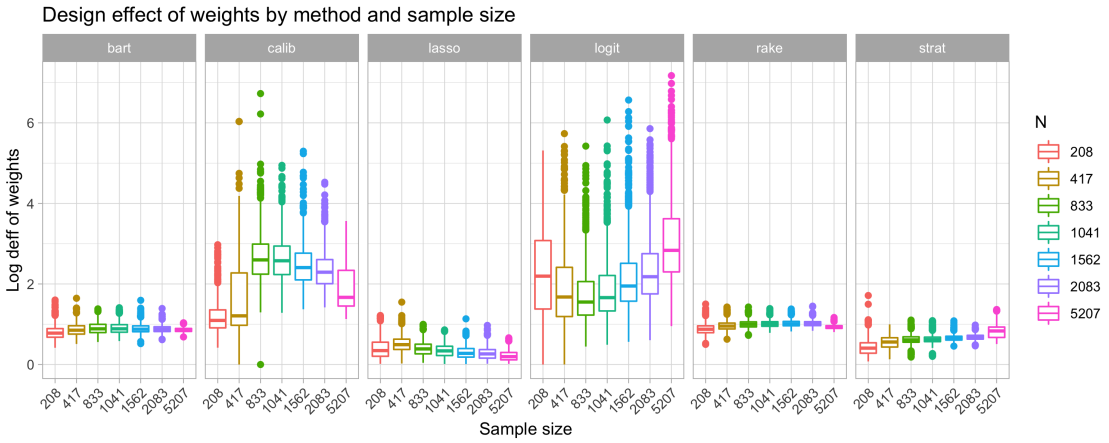
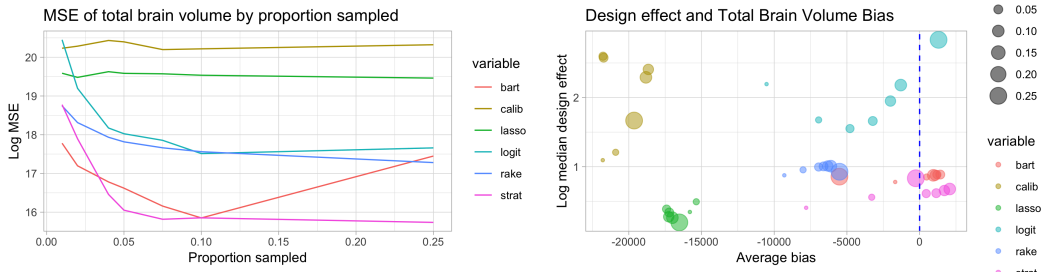


Figure 5.7: Log design effect by method and sample size in simulation studies

similarly, with a sharp drop in MSE as the proportion sampled increases until 5%, then a steady decrease. Calibration and the LASSO have consistent, high MSE compared to the other methods.

Bias and design effect The ideal estimator will eliminate bias without a introducing a large design effect. Figure 5.8b shows the relationship between the log median bias and the log median design effect across methods (shown in color) and proportion sampled (size of each point). We use log medians due to the large skew in the distributions of each variable. The LASSO estimator has the smallest design effect, but hardly eliminates bias. Conversely, the logit effectively eliminates bias as proportion sampled increases, but at the cost of a large design effect. Calibration is clearly the worst performing, as it fails to reduce bias and also has a large design effect. BART and post-stratification both effectively eliminate bias without a large design effect. BART has a small advantage at the smaller sample sizes, but post-stratification quickly catches up.



(a) MSE of total brain volume by proportion sampled

(b) Log median design effect of weights by average bias in total brain volume

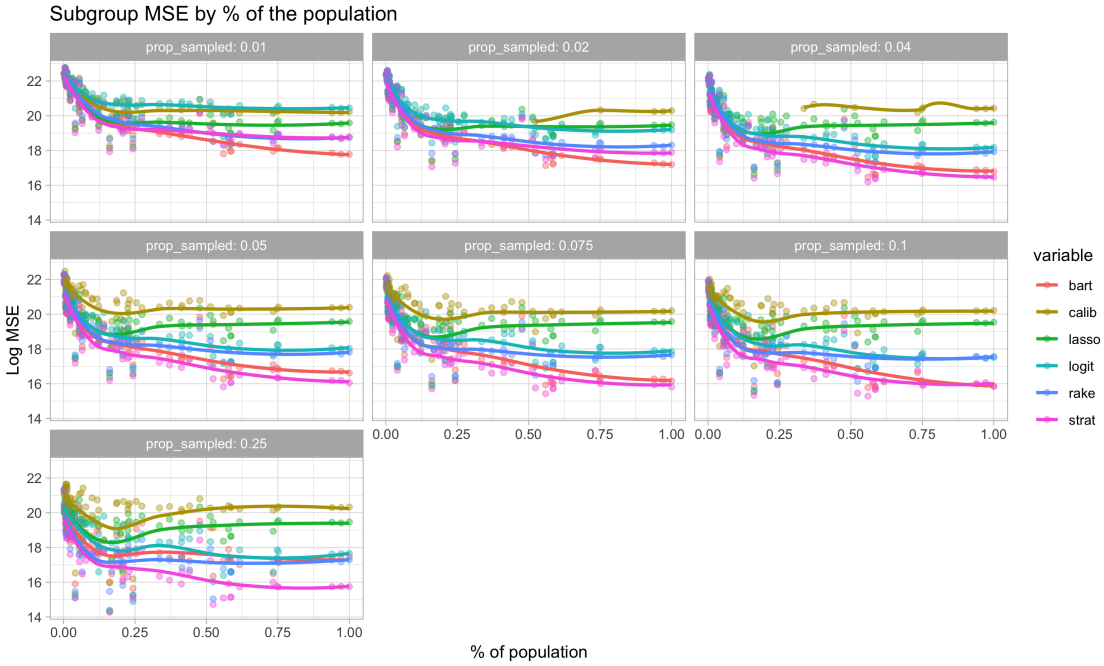


Figure 5.9: Points represent the log MSE of total brain volume calculated for demographic subgroups, by the true proportion of the population made up by that subgroup. Lines are smoothed estimates of the association between proportion of the population and MSE.

Subgroup estimation Trends from the metrics discussed thus far persist at the subgroup level. Figure 5.9 shows log MSE of total brain volume within key subgroups as a function of the true size of the subgroup in the population. There is no clear difference in performance between methods in the smallest subsets, however BART has the lowest MSE for larger subgroups in the two smallest samples, slightly outperforming raking and post-stratification. Post-stratification outperforms other methods in larger subgroups as sampled proportion increases. Calibration consistently has the highest MSE, with the LASSO not far behind.

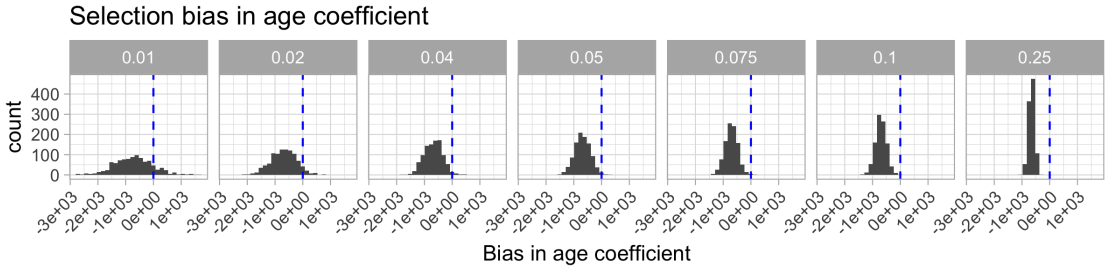


Figure 5.10: The distribution of estimates of the coefficient for age in a linear regression of total brain volume on age. Regressions were estimated using unadjusted (selection-biased) sample data. The blue lines represent the true population value of the age coefficient.

Age and total brain volume

For each sample, we regress total brain volume on age, once in a simple linear regression, and once in a weighted linear regression for each adjustment method. We record the estimated values of the intercept and age coefficient. Figure 5.10 shows the simulated selection bias in the age coefficient from these regressions, estimated with the unweighted linear regression. The histograms are distributions of the age coefficient, and the blue dotted line represents population truth. We can see that there is a small amount of selection bias in the unadjusted age coefficients.

Figure 5.11 shows the bias in the age coefficient remaining after adjustment procedures were applied. While it appears that all procedures are relatively unbiased, this is more a function of the small amount of selection bias introduced (as a proportion of the total variation in the age coefficient) rather than performance of the methods. The logit approach seems to produce the only truly unbiased estimator of the age coefficient, however we can see in Figure 5.12 that the MSE of the logit method is still quite high. Though BART and post-stratification don't seem to completely eliminate the selection bias in the estimate of the association between total brain volume and age, they do so without introducing a large amount of variance, so on the whole seem to perform slightly better than other methods.

Population composition

The last metric by which we evaluate adjustment methods is distribution bias (DB). Figure 5.13 shows the distribution of DB across simulations for each adjustment method by sample size. Raking has the smallest DB across all sample sizes considered, and the DB decreased as sample size grew. DB for the logit method was the most variable, but, on average, decreased the most drastically with increases in sample size. DB for both post-stratification and BART was less variable across

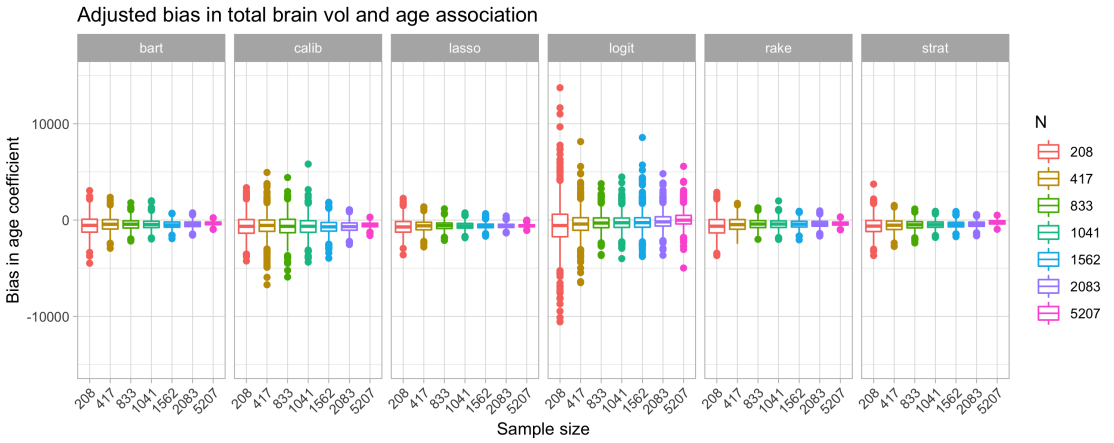


Figure 5.11: Bias in weighted estimate of age coefficient in regression predicting total brain volume.

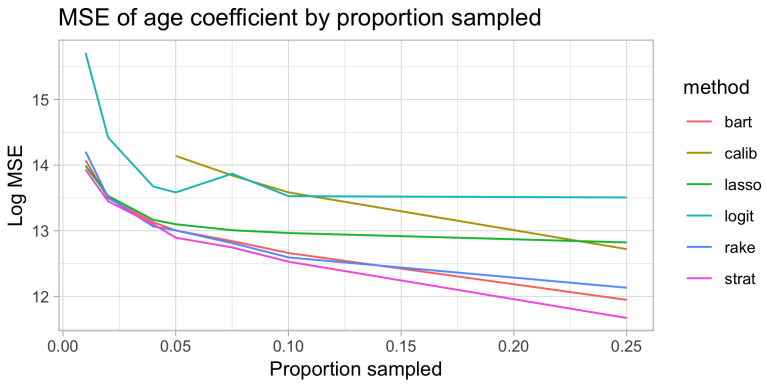


Figure 5.12: Log MSE as a function of proportion sampled.

simulation iterations, and consistent across sample sizes, but higher on average in larger sample sizes than raking or logit.

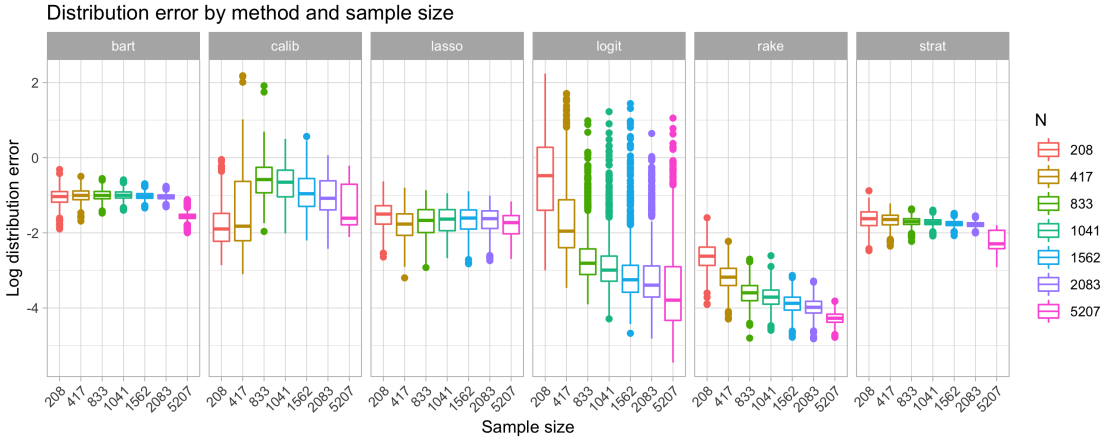


Figure 5.13: Log distribution bias by method adjustment method and sample size.

5.6.3 Application to the UK Biobank

Table 5.3 gives adjusted estimates of prevalence for selected health outcomes from the UK Biobank imaging cohort, related to unadjusted UK Biobank data and population estimates from the HSE. Weighting generally seems to improve prevalence estimates for most health outcomes.

Take, for example, the proportion of the population estimated to be current smokers. The HSE estimates that 15.1% of the population smokes, while only 3.9% of the subjects in the UK Biobank imaging cohort report being current smokers. All methods of adjustment improve this estimate. BART and raking estimate that 7.4% and 7.6% of the population currently smokes, while the lowest estimate is from the LASSO which estimates that 4.6% of the population smokes.

Similarly, for obesity, BART estimates that 24.2% of the population is obese, compared to 28.5% in the HSE, while the unadjusted estimate is only 18.7%. Other methods improve on the unadjusted estimate, however none so much as BART. Stratification also seems to drastically improve estimates of these health quantities, while other methods, like the LASSO and logit, seem to lag behind. It is important to note that while these estimators improve on the unadjusted estimator, even the best-performing are not able to completely eliminate selection bias.

Table 5.4 gives weighted estimates of total brain volume relative to the unweighted estimate from the UK Biobank imaging cohort. The weighted estimates have hardly changed from the unweighted estimate (the largest difference is from BART, which changes only by about 1% of the unweighted total brain volume estimate).

Table 5.5 shows the design effect, distribution bias and estimated age coefficient for each method. The results are in-line with simulation studies. Raking is the best at matching population totals of auxiliary variables (as measured by the lowest DB), with BART and post-stratification performing almost as well. Calibration matching population totals reasonably well, but has the largest design effect. The logit approach does not reduce distribution bias from the unweighted estimate, which is consistent with the known shortcomings of the method. BART, as a regression-based approach, improves considerably on the logit baseline.

Age coefficients are adjusted down from the unweighted estimator, but not to a significant degree. There is not much differentiation in estimated association between age and total brain volume across adjustment methods.

Table 5.3: Weighted estimates of prevalence health outcomes from the UK Biobank imaging cohort data, compared to our target population, the HSE.

			UK Biobank adjusted (%)					
Level	HSE %	UKB Raw %	Rake	Post-strat	Calib	LASSO	Logit	BART
ApoE Phenotype								
01-e4/e4	-	2.2	1.7	1.9	1.8	2.1	2.2	1.8
02-e3/e4	-	23	23.3	23.4	23.1	22.9	23	23.5
03-other	-	74.8	74.9	74.7	75	75	74.8	74.6
BMI Bucket								
01-Underweight	0.6	0.7	0.7	0.8	0.5	0.7	0.7	0.7
02-Healthy	24.6	37	32.6	31.9	32.9	36	36.7	33
03-Overweight	37.9	41.2	40.7	41.5	41.6	41.5	41.2	40.2
04-Obese	28.5	18.7	24.1	23.6	22.9	19.4	19.1	24.2
99-DNK/Refused	8.6	2.4	1.9	2.2	2.1	2.3	2.3	1.9
Diabetes Ever								
01-Yes	10.4	5.1	6.5	7.1	5.5	5.2	5.3	6.3
02-No	89.6	94.9	93.5	92.9	94.5	94.8	94.7	93.7
99-DNK/Refused	0.1	-	-	-	-	-	-	-
High BP Ever								
01-Underweight	0.6	0.7	0.7	0.8	0.5	0.7	0.7	0.7
02-Healthy	24.6	37	32.6	31.9	32.9	36	36.7	33
03-Overweight	37.9	41.2	40.7	41.5	41.6	41.5	41.2	40.2
04-Obese	28.5	18.7	24.1	23.6	22.9	19.4	19.1	24.2
99-DNK/Refused	8.6	2.4	1.9	2.2	2.1	2.3	2.3	1.9
Smoking Status								
01-Current	15.1	3.9	7.6	6.2	6.9	4.6	4.1	7.4
02-Previous	36	33.6	32.3	34.7	31.1	32.8	33.5	33.1
03-Never	48.8	61.6	59.6	57.9	59.6	61	61.4	59.1
99-DNK/Refused	-	1	0.5	1.2	2.4	1.6	1	0.4

5.7 Discussion

5.7.1 Simulation studies

The simulation studies revealed BART and post-stratification to be the most effective at reducing bias without introducing much additional variance. BART slightly outperformed post-stratification in the smallest samples, likely the more realistic scenarios than observing 25% of the target population. BART, however, only did a mediocre job of matching sample marginal distributions to those of the population. This is likely due to the fact that raking variables were selected based on which were the most important predictors of selection, and not based on which represented the largest subgroups in the population. Altering this selection criterion could improve BART's performance along this metric.

Calibration performed almost remarkably poorly, failing to reduce bias while drastically increasing variance compared to other methods. It performed far worse

Table 5.4: Adjusted estimates of total brain volume from the UK Biobank imaging cohort

Method	Distribution Bias	deff	β_{age}
Unadjusted	0.57	1	-5316
BART	0.06	5.85	-5018
Calib	0.11	7.89	-4915
LASSO	0.29	1.78	-5047
Logit	0.49	1.03	-5269
Rake	0.01	6.77	-5092
Post-strat	0.10	4.8	-5252

Table 5.5: The distribution bias (DB) and design effect (deff) of each adjustment method relative to unadjusted estimates.

Method	Distribution Bias	deff
Unadjusted	0.57	1
BART	0.06	5.85
Calib	0.11	7.89
LASSO	0.29	1.78
Logit	0.49	1.03
Rake	0.01	6.77
Post-strat	0.10	4.8

than raking, to which its closely related, except that calibration included constraints based on the continuous form of age, while raking relied on only the discrete specification of age. It is possible that since we only calculated performance of methods using categorical forms of variables, our assessment did not fully capture advantages of calibration. This could also be due to poor specification of algorithm parameters, and warrants further examination.

That LASSO performed poorly was also surprising, as it is also so closely related to raking. It may be that the LASSO is not able to reliably select the most important variables for adjustment, or that our method of creating tiers of variables for sequential raking is not the optimal strategy for handling a large number of selected variables.

From an implementation standpoint, BART has two main advantages over other methods: variable selection is done automatically, and complex interactions are implicitly considered without the need to enumerate all of them. LASSO and logit include a mechanism for variable selection, but seem to lack the power that BART has to identify critical variables in small samples. They also both select interactions only from a set previously specified by the researcher, which limits the degree of interactions that can be considered before hitting computation time and memory errors. Post-stratification, in the way we have implemented it here, also

considers interactions and has an automatic variable selection feature, likely why it performs similarly to BART. However, in small samples, there is a limit to the number of variables that can be used for post-stratification before strata become too small. BART has no such limitation, likely why it outperforms post-stratification in smaller samples. In large samples, post-stratification can consider high-degrees of interactions, and will dominate performance.

Logit adjustment is the only method without specific constraints of matching weighted sample marginal distributions to population distributions, however, in small samples had a DB on par with other methods, on average, and in larger samples had some of the lowest DB of any method. One caveat is that though the DB was low on average, it was highly variable.

It is important to note that most of the methods that we consider here require access not only to external population data, but specifically to individual-level population data, which is not always available. Our best-performing methods, BART and post-stratification, are not possible without access at least to joint distributions of auxiliary variables. Other methods, like raking and calibration, though they perform worse in this setting are still highly useful in other scenarios.

5.7.2 UK Biobank application

In the application of these methods to real UK Biobank data, we observed that they were able to improve estimates of prevalence of smoking, diabetes, obesity and high blood pressure relative to population estimates from the HSE. BART and post-stratification exhibited the largest improvements over the unweighted estimators, though were still unable to eliminate selection bias completely.

5.7.3 Limitations

There are numerous caveats and limitations of the results presented here. First, the data we use as the target population is itself a study based on a population sample, so the population quantities that we treat as true are in fact uncertain. Unusually, the target population is also much smaller that the UK Biobank imaging cohort, adding additional uncertainty to the analysis. Furthermore, bias that we attribute here to preferential selection may be from another source, like measurement error. For example, the HSE and UK Biobank have different questionnaires, and, for example as respondents to report education level in slightly different ways. While care has been taken here to standardize the responses across sources, there is likely some lingering discrepancy.

Second, the crucial assumption underlying all adjustment procedures tested here is that we have correctly identified an admissible set. While we introduce the concept of an admissible set and criteria necessary for recovering from selection bias, we do not actually identify such a set in the UK Biobank. We also limit our analysis only to auxiliary variables for which we have external population data readily available, when that is not a requirement for recovery in all cases. Furthermore, there is a wealth of data available in the UK Biobank that we fail to leverage, perhaps most obviously spatial data.

The third major limitation involves our treatment, or lack thereof, of variance of estimators or statistical significance of results. We only evaluate relative performance based on visual assessments, not based on statistical tests, so cannot make any claims about a method having significantly better performance than another.

Lastly, we only perform simulation studies in which the missingness mechanism is static. Adjustment methods that performed well could be particularly suited to the characteristics of the missingness used here.

5.7.4 Future Research

There are many possible avenues for future research. First, we could continue to explore and refine various adjustment procedures. For example, we could adapt calibration and raking to better handle larger numbers of variables. We could also expand our analysis to consider methods that directly model the outcome of interest, like MRP. We could explore tuning parameters for the BART and LASSOs.

Second, we could incorporate additional auxiliary variables, like spatial data, into weighting procedures. As discussed, many weighting procedures suffer from an inability to handle a large number of auxiliary variables, requiring the researcher to manually select variables and interactions they think will be important

Another direction for future research is incorporating computation time into overall analysis of estimators. BART, for example, performs well, but also takes exorbitantly more computation time than post-stratification, for example.

5.8 Supplementary Material

All code is publicly available at https://github.com/vcbradley/ukb-selection-bias.

Table 5.6: Missingness coefficients used in the simulation study.

Z	β_z
age	2.2362584
demo_sexMale	-2.3681210
demo_age_bucket60 to 64	-0.5221229
demo_ethnicity_full99-DNK/Refused	-0.3310816
demo_empl_retired	-0.6548054
demo_empl_unemployed	1.0396497
demo_empl_student	-1.2736848
demo_occupation04-admin	1.0748084
demo_occupation99-DNK/Refused	-0.6608140
demo_educ_collegeplus	-0.7038866
demo_educ_highest_full02-A Levels	1.8571014
demo_educ_highest_full03-O Levels	-0.6431069
demo_educ_highest_full07-None	0.5751345
demo_income_bucket04-52k to 100k	2.3920356
demo_hh_size4	0.5362810
demo_hh_ownrent02-Own with mortgage	-0.6211912
demo_hh_ownrent99-DNK/Refused	0.8182660
age_sq	0.0915620
age:demo_ethnicity_full04-Mixed	-1.6169544
demo_age_bucket50 to 54:demo_empl_disabled	-0.7442275
demo_ethnicity_full05-Asian Indian:demo_empl_volunteer	2.0003787
demo_ethnicity_full06-Asian Bangladeshi:demo_hh_ownrent06-Rent free	-2.2941626
demo_empl_retired:demo_hh_ownrent06-Rent free	-1.1010164
demo_empl_disabled:demo_occupation07-sales customer service	0.7285372
demo_empl_disabled:demo_hh_ownrent03-Rent from LA	1.5001765
demo_occupation08-industrial:demo_hh_ownrent04-Rent private	1.7046601

5.9 Acknowledgments

Computation used the BMRC facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. Additionally, the author would like to thank Tom Nichols for supervision and incredibly helpful guidance.

5.10 Figures and Tables

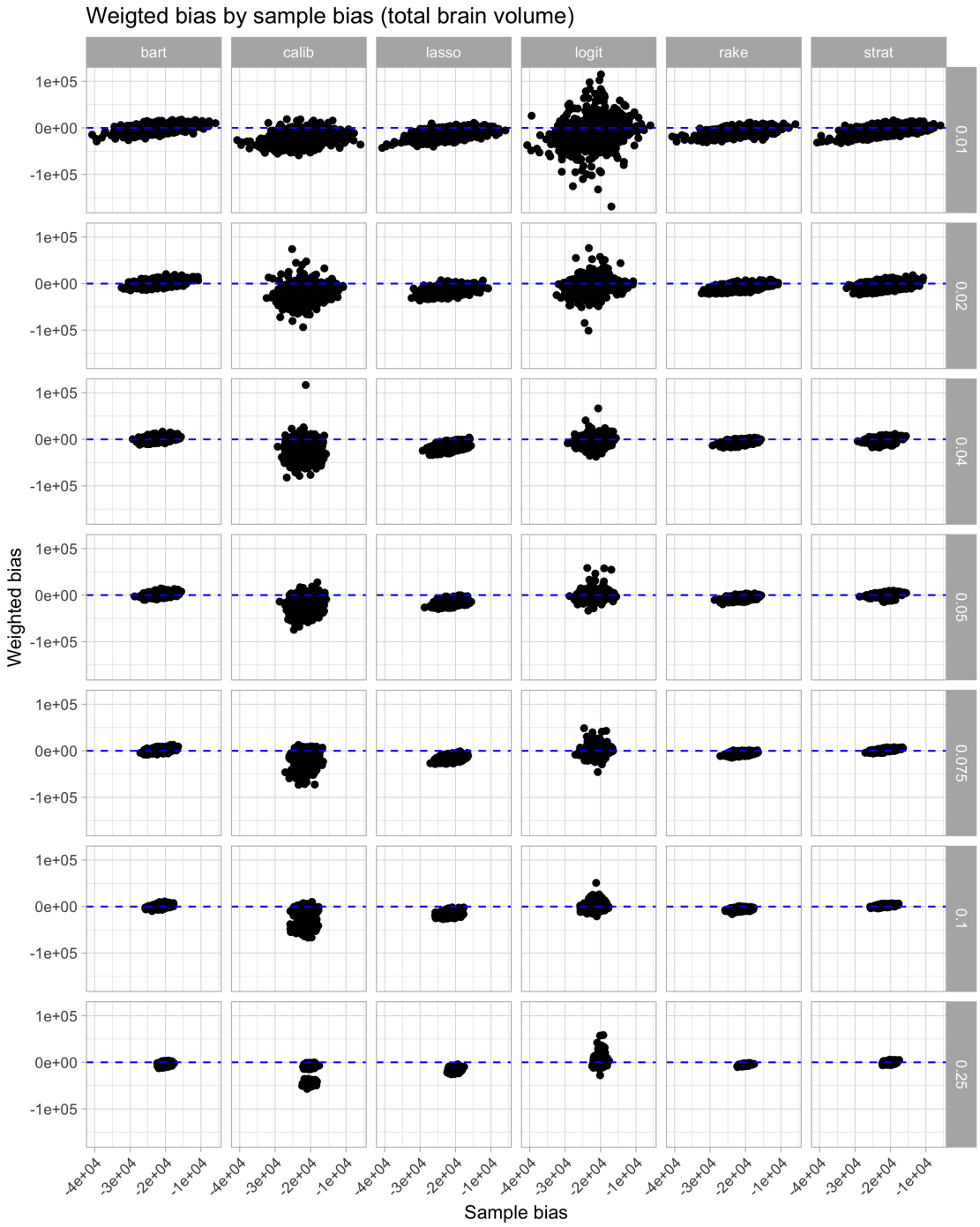


Figure 5.14: Bias in total brain volume. The x-axis shows the actual selection bias in the sample, the y-axis shows the remaining bias once each adjustment procedure was applied.

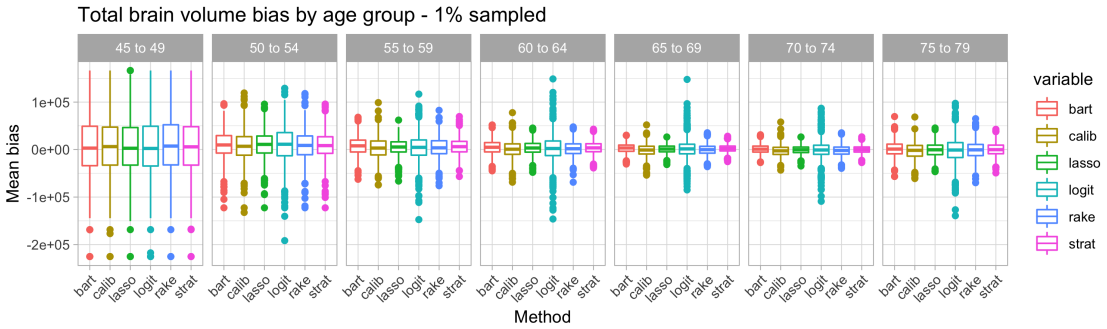


Figure 5.15: Bias of total brain volume by proportion sampled

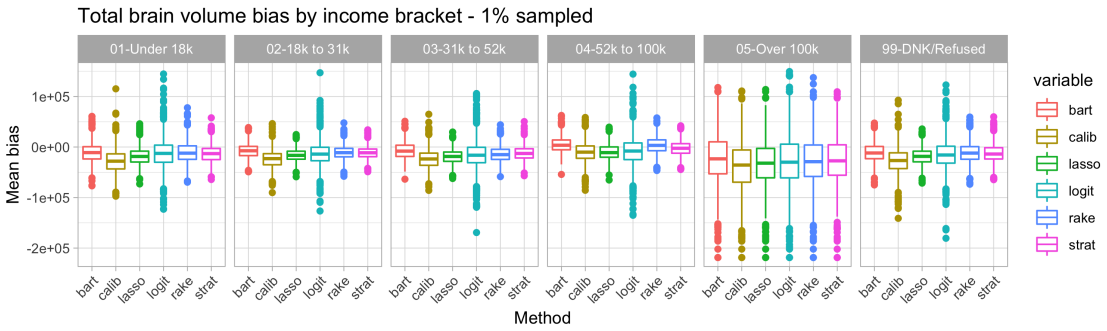


Figure 5.16: Log median design effect of weights by average bias in total brain volume

Algorithm 2: Algorithm for second proposed simulation where missingness is re-randomized on each iteration.

Result: Weighted samples

for $\pi_{obs} \in (0.01, 0.02, 0.04, 0.05, 0.075, 0.1, 0.25, 0.5)$ **do**
  $n_{sim} = \pi_{obs} * N$;
  for $m \in (1, \ldots 1000)$ **do**
    sample missingness coefficients $\boldsymbol{\beta}$;
    calculate probability of missingness p_i for all N subjects $p_i = \text{logit}(\mathbf{X}_i\boldsymbol{\beta})$;
    select sample of n_{sim} subjects;
    $s_i = 1$ if i^{th} subject is selected where $s_i \sim \text{Bern}(p_i|n_{sim})$;
    for *each adjustment procedure* **do**
      weight sample;
      return *weights*
    end
  end
end

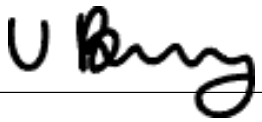
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Methods for selection bias adjustment in the UK Biobank neurological imaging data
Publication Status	☐Published ☐ Accepted for Publication ☐Submitted for Publication ☒Unpublished and unsubmitted work written in a manuscript style
Publication Details	**Addressing selection bias in the UK Biobank neurological imaging cohort** **Valerie Bradley, Thomas E. Nichols** **medRxiv 2022.01.13.22269266; doi: https://doi.org/10.1101/2022.01.13.22269266**

Student Confirmation

Student Name:	Valerie C Bradley		
Contribution to the Paper	Jointly developed the idea with T.N.; performed all statistical analysis; wrote manuscript		
Signature 		Date	Jan 5, 2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: **Prof Dino Sejdinovic**		
Signature 	Date	7 January 2024

This completed form should be included in the thesis, at the end of the relevant chapter.

6
Conclusion

Selection bias presents a challenge to any researcher interested in making population inferences using a sample from the population. At the same time, selection bias is nearly impossible to prevent, difficult to measure, dynamic even within a particular context, requires luck to recover from, and can have a catastrophic impact in practical applications. In other words, a problem worth endless study.

Minimizing the impact of selection bias on population inferences requires more than an optimistic reliance on probability sampling theory, but rather a commitment to total selection bias minimization. Researchers must consider selection bias in every step of the research process – from design, to fieldwork, to analysis, and communication. This thesis examines selection bias from a variety of perspectives and aims to enrich the set of tools available to researchers for preventing, quantifying, adjusting for, and communicating about selection bias.

Chapter 2 focuses on the quantification of selection bias that remains in surveys after data is collected and persists even once best efforts are made to adjust for selection bias. We demonstrate how the *ddc* framework from Meng (2018) can be used to estimate the contribution that selection bias makes to overall error, and, critically, the mathematical inefficiency of attempting to compensate for low data quality by increasing data quantity. We also highlight how selection bias may be *dynamic* over time, even holding constant a data collection procedure and outcome, and show that selection bias can result in over a 99% reduction in bias-adjusted effective sample size, rendering population estimates from a survey of over 250,000 respondents no more precise than a survey of only 1,000 respondents.

However a more subtle finding from this paper is the one that has perhaps had the largest impact on the rest of my research. We introduce how *ddc* can be used to

decompose error in the estimate of a population mean by *stage* of data collection. Meng (2018) introduces the *Law of Large Populations*, explaining how the impact of selection bias is moderated by population size. We expand on this finding and show, mathematically, that selection bias in each *stage* of data collection is similarly moderated by the relevant population size at that stage. Therefore, as the relevant population size decreases at each stage of data collection, the earliest stage of data collection at which selection bias is first allowed to affect the response mechanism is the stage that will dominate the overall *ddc* because it has the largest relevant population size. The *ddc* in subsequent stages may attempt to cancel out earlier *ddc*, but the relative population sizes make this a challenging task. In practical terms, this suggests that a probabilistic sampling mechanism does little to overcome any selection bias that has already been introduced upstream into the sampling frame.

This implication of the *ddc* framework pushed the focus of my research away from nonresponse adjustment methods (e.g. those in Chapter 5), which can only attempt to recover from selection bias that has already been observed, and instead towards sampling mechanisms that attempt to anticipate and preempt selection bias (ALSD in Chapter 3), and definition of and uncertainty in the population of interest itself (Leverage in Chapter 4).

This work only scratches the surface of what can and must be done to tackle selection bias. First, there are a number of limitations to each of the approaches and methods presented here. Though a powerful theoretical framework, using *ddc* in practice generally requires that ground truth has been observed, which is rare in real world settings. For example, pre-election polls which seek to measure support for political candidates may be compared against election day outcomes, but final candidate vote share may not actually reflect true candidate support when a poll was administered. Outside of pre-election polling, it is extremely rare to observe population outcomes for quantities measured with public opinion polling. While *ddc* may be useful for retroactively diagnosing selection bias that has already occurred, it does not reveal *why*. How best to use *ddc* to proactively prevent selection bias or how well *ddc* generalizes across different nonresponse settings remains an open question.

ALSD presents a flexible and promising framework for how to design samples to meet modern analysis needs, but has not yet been fully explored. In the paper presented here, we focus mainly on the multi-outcome setting which is most common in survey research, rather than the single-outcome setting which is more closely related to AL and BO. Our current implementation of single-outcome ALSD does not appear to be competitive with the multi-outcome implementation, even when assessing the very outcome that the single-outcome setting has been designed to

estimate. We believe this will improve with more fine-tuning of the implementation, for example, by using a more flexible outcome model. Second, as we have discussed throughout this work, selection bias is dynamic. While ALSD is well-positioned to adapt to such dynamic selection bias, the implementation presented here gives equal weight to all prior observations, so as more data is collected, the modeled response propensity may be slow to update to new response environments. An interesting extension of ALSD would involve weighting observations by recency to allow modeled response propensity to adapt more efficiently.

As discussed in the Introduction, the contactability landscape for survey research becomes more fractured with each passing year. One core challenge in survey design is allocating target responses to potential modes of outreach, yet to date, guidance on how best to perform this allocation is sparse. The ALSD framework can be extended to incorporate multiple modes of outreach, each with different cost structures, to help researchers design samples that are efficient both with respect to sample size and budget, and, critically, targeting respondents where they are most contactable.

Our immediate next step with ALSD is to examine the results of the live test recently conducted with the WFP in Zimbabwe. We collected 1500 responses over four waves using ALSD, alongside the WFP's normal data collection program for their mVAM initiative. This will allow us to assess under real-world conditions the efficiency of ALSD relative to widely-used quota sampling techniques.

Though the notion of leverage that we develop in Chapter 4 tackles selection bias at the analysis stage, it takes a different perspective than most nonresponse adjustment methods. Most nonresponse adjustment methods ignore any uncertainty in population benchmarks used in adjustment, and thus produce overly-confident population inferences. Leverage does not aim to fix selection bias at this stage so much as help the researcher present findings with a more comprehensive assessment of total uncertainty. Simulations are useful tools for exploring this type of epistemic uncertainty, but can be unwieldy, and leverage and associated population uncertainty intervals aim to provide a simple and easy tool to aid analysis. However, there are a few key limitations to leverage in its current form. First, leverage is currently only defined in the case of a binary population target, and second, population uncertainty intervals only allow researchers to consider the uncertainty in a single population target at a time. In the future, we hope to address these limitations.

Throughout this work I hope to have communicated a sense of urgency around the need to address selection bias head-on in any setting in which researchers strive to make population inferences using data collected from only a subset of the population, whether that subset is a near-census collected by a government agency, or

a survey that uses a probabilistic sampling mechanism but suffers from nonresponse. High-quality inferences require high-quality data, and high-quality data requires the prevention, quantification, mitigation, and communication of selection bias.

Bibliography

C. Appel, D. Beltekian, D. Gavrilov, C. Giattino, J. Hasell, B. Macdonald, E. Mathieu, E. Ortiz-Ospina, H. Ritchie, L. Rodés-Guirao, and M. Roser. Data on COVID-19 (coronavirus) by Our World iin Data, 2021. URL `https://github.com/owid/covid-19-data`.

A. Arrieta, E. Gakidou, H. Larson, E. Mullany, and C. Troeger. Through Understanding and Empathy, We Can Convince Women to Get the COVID-19 Vaccine. *Think Global Health*, 2021.

Associated Press-NORC Center for Public Affairs Research. The June 2021 AP-NORC Center Poll. 2021. URL `https://perma.cc/6ZXM-58XT`.

B. Auxier and M. Anderson. Social Media Use in 2021. *Pew Research Center*, 2021.

M. A. Bailey. A New Paradigm for Polling. *Harvard Data Science Review*, 5(3), 2023. URL `https://hdsr.mitpress.mit.edu/pub/ejk5yhgv`.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

E. Bareinboim, J. Tian, and J. Pearl. Recovering from selection bias in causal and statistical inference. *Proceedings of The Twenty-Eighth Conference on Artificial Intelligence*, (July):339–341, 2014.

N. Barkay, C. Cobb, R. Eilat, T. Galili, D. Haimovich, S. Larocca, K. Morris, and T. Sarig. Weights and methodology brief for the COVID-19 Symptom Survey by University of Maryland and Carnegie Mellon University, in partnership with Facebook. 2020. URL `https://arxiv.org/abs/2009.14675`.

S. Benonisdottir and A. Kong. Studying the genetics of participation using footprints left on the ascertained genotypes. *Nature Genetics*, 55(8):1413–1420, 2023.

A. J. Berinsky. American public opinion in the 1930s and 1940s: The analysis of quota-controlled sample survey data. *International Journal of Public Opinion Quarterly*, 70(4):499–529, 2006.

J. Bethlehem. Weighting Nonresponse Adjustments based on Auxiliary Information. In R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, editors, *Survey Nonresponse*, pages 275—-288. Wiley, New York, 2002.

J. G. Bethlehem and B. Schouten. *Nonresponse adjustment in household surveys.* Statistics Netherlands Voorburg/Heerlen, The Netherlands, 2004.

P. P. Biemer and L. E. Lyberg. *Introduction to survey quality.* John Wiley & Sons, 2003.

S. J. Blumberg and J. V. Luke. Wireless Substitution: Early Release of Estimates Based on Data from the National Health Interview Survey, July-December 2006. *National Center for Health Statistics*, 2007.

S. J. Blumberg and J. V. Luke. Wireless Substitution: Early Release of Estimates From the National Health Interview Survey, July-December 2022. *National Center for Health Statistics*, 2022.

M. Blumenthal. Why YouGov is changing how we ask people whether they've received the COVID-19 vaccine. 2021. URL `https://perma.cc/2EYN-K358`.

P. Bouman, V. Dukic, and X.-L. Meng. A Bayesian multiresolution hazard model with application to an AIDS reporting delay study. *Statistica Sinica*, pages 325–357, 2005.

V. C. Bradley and T. E. Nichols. Addressing selection bias in the UK Biobank neurological imaging cohort. *medRxiv*, 2022. doi: `https://doi.org/10.1101/2022.01.13.22269266`.

V. C. Bradley, S. Kuriwaki, M. Isakov, D. Sejdinovic, X.-L. Meng, and S. Flaxman. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890):695–700, 2021.

J. A. Brandon, B. C. Farmer, H. C. Williams, and L. A. Johnson. APOE and Alzheimer's disease: Neuroimaging of Metabolic and Cerebrovascular Dysfunction. *Frontiers in Aging Neuroscience*, 10(JUN):1–8, 2018.

N. Breznau, E. M. Rinke, A. Wuttke, H. H. Nguyen, et al. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, 119(44):e2203150119, 2022.

British Polling Council. Performance of the Polls in the EU Referendum, 2016. URL `https://www.britishpollingcouncil.org/performance-of-the-polls-in-the-eu-referendum/`.

L. D. Brown, T. T. Cai, and A. DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16:101–117, 2001.

H. Carpenter. Understanding Society: Waves 1-10, 2009-2019 and Harmonised BHPS: Waves 1-18, 1991-2009. Wave 8 technical report. 2018. URL `https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/mainstage/technical-reports/wave-8-technical-report.pdf`.

D. Caughey and E. Hartman. Target Selection as Variable Selection : Using the Lasso to Select Auxiliary Vectors for the Construction of Survey Weights. *Available at SSRN 3494436*, pages 1–33, 2017.

R. Chenevert, A. Gottschalck, M. Klee, and X. Zhang. Where the wealth is: the geographic distribution of wealth in the united states. *US Census Bureau*, 2017.

S. Clement, D. Blaz, and E. Guskin. Post-ABC polls: Biden leads Trump narrowly in Michigan, significantly in Wisconsin, 2020. URL `https://www.washingtonpost.com/politics/2020/10/28/wisconsin-michigan-poll-post-abc/?itid=sf_politics-polling`.

J. Clinton, J. Agiesta, M. Brennan, C. Burge, M. Connelly, A. Edwards-Levy, B. Fraga, E. Guskin, D. S. Hillygus, C. Jackson, et al. Pre-Election Polling: An Evaluation of the 2020 General Election Polls. *American Association for Public Opinion Research*, 2020.

S. Coffey, B. T. West, J. Wagner, and M. R. Elliott. What do you think? Using expert opinion to improve predictions of response propensity under a Bayesian framework. *Methods, data, analyses*, 14(2), 2020.

S. M. Coffey. *Bayesian Methods for Prediction of Survey Data Collection Parameters in Adaptive and Responsive Designs*. PhD thesis, University of Maryland, College Park, 2020.

N. Cohn, 2020. Tweet. 29 October 2020. URL:`https://twitter.com/Nate_Cohn/status/1321833426870325254?s=20`.

J. D. Correa, J. Tian, and E. Bareinboim. Generalized adjustment under confounding and selection biases. *32nd AAAI Conference on Artificial Intelligence*, (June): 6335–6342, 2018.

F. R. Day, P. R. Loh, R. A. Scott, K. K. Ong, and J. R. Perry. A Robust Example of Collider Bias in a Genetic Association Study. *American Journal of Human Genetics*, 98(2):392–393, 2016.

Democracy Docket. Revisiting Polling for 2021 and Beyond, 2021. URL `https://www.democracydocket.com/2021/04/revisiting-polling-for-2021-and-beyond/`.

W. Dempsey. The hypothesis of testing: Paradoxes arising out of reported Coronavirus case-counts. 2020. URL `https://arxiv.org/abs/2005.10425`.

A. J.-C. Deville and C.-E. Särndal. Calibration Estimators in Survey Sampling Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992.

J.-C. Deville, C.-E. Särndal, and O. Sautory. Generalized Raking Procedures in Survey Sampling Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 1459(January):1013–1020, 1993.

N. Egami and E. Hartman. Covariate selection for generalizing experimental results: application to a large-scale development program in Uganda. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(4):1524–1548, 2021.

M. R. Elliott and R. Valliant. Inference for nonprobability samples. *Statistical Science*, 32(2):249–264, 2017.

R. E. Fay and G. Train. Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties. *Proceedings of the Section on Government Statistics, American Statistical Association*, pages 154–159, 1995.

J. F. Fields and J. Hunter-Childs *et al.* Design and operation of the 2020 Household Pulse survey. 2020. U.S. Census Bureau. `https://perma.cc/JC3D-3LBY`.

A. F. Fotenos, M. A. Mintun, A. Z. Snyder, J. C. Morris, and R. L. Buckner. Brain Volume Decline in Aging. *Archives of Neurology*, 65(1):113–120, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 2010.

A. Fry, T. J. Littlejohns, C. Sudlow, N. Doherty, L. Adamska, T. Sprosen, R. Collins, and N. E. Allen. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American Journal of Epidemiology*, 186(9):1026–1034, 2017.

S. K. Gadarian, S. W. Goodman, and T. B. Pepinsky. Partisanship, health behavior, and policy attitudes in the early stages of the COVID-19 pandemic. *PLOS One*, 16(4), 2021.

Gallup. How Did the Gallup-Sharecare Well-Being Index Work? 2018. URL `https://www.gallup.com/224870/gallup-sharecare-index-work.aspx`.

Y. Gao, L. Kennedy, D. Simpson, and A. Gelman. Improving multilevel regression and poststratification with structured priors. *Bayesian Analysis*, 16(3):719, 2021.

R. Garnett. *Bayesian optimization*. Cambridge University Press, 2023.

A. Gelman. Comment: Struggles with Survey Weighting and Regression Modeling. *Statistical Science*, 22(2):153–164, 2007.

A. Gelman. Reverse-engineering the problematic tail behavior of the Fivethirtyeight presidential election forecast, 2020. URL `https://statmodeling.stat.columbia.edu/2020/10/24/reverse-engineering-the-problematic-tail-behavior-of-the-fivethirtyeight-presidential-`

A. Gelman. Failure and success in political polling and election forecasting. *Statistics and Public Policy*, 8(1):67–72, 2021.

A. Gelman and T. C. Little. Poststratifcation Into Many Categories Using Hierarchical Logistic Regression. *Survey Methodology*, 1997:1–26, 1997.

A. Gelman and A. Vehtari. What are the most important statistical ideas of the past 50 years? *Journal of the American Statistical Association*, 116(536):2087–2097, 2021.

A. Gelman, S. Goel, D. Rivers, and D. Rothschild. The Mythical Swing Voter. *Quarterly Journal of Political Science*, 11(1):103–130, 2016.

L. Geuzinge, J. van Rooijen, and B. Bakker. The use of administrative registers to reduce non-response bias in household surveys. *Netherlands Official Statistics*, 15:32–38, 2000.

Y. Ghitza and A. Gelman. Voter registration databases and MRP: Toward the Use of large-scale databases in public opinion research. *Political Analysis*, 28(4):507–531, 2020.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate Shift by Kernel Mean Matching. In J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, editors, *Dataset Shift in Machine Learning*, pages 131–160. 2013.

J. Groen. Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. *Journal of Official Statistics*, 28:173–198, 2012.

R. M. Groves. Three eras of survey research. *Public Opinion Quarterly*, 75(5 SPEC. ISSUE):861–871, 2011.

R. M. Groves and S. G. Heeringa. Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):439–457, 2006.

R. M. Groves and L. Lyberg. Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5):849–879, 2010.

R. M. Groves and E. Peytcheva. The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, 72(2):167–189, 2008.

R. M. Groves, F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey methodology*. John Wiley & Sons, 2011.

E. J. Haas, F. J. Angulo, J. M. McLaughlin, E. Anis, S. R. Singer, F. Khan, N. Brooks, M. Smaja, G. Mircus, K. Pan, et al. Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *The Lancet*, 2021.

E. Hartman. "Methods to Alleviate Unit Non-Response Bias in Surveys". unpublished, 2014.

E. Hartman and M. Huang. Sensitivity analysis for survey weights. *Political Analysis*, 32 (1):1–16, 2024.

E. Hartman, C. Hazlett, and C. Sterbenz. kpop: A kernel balancing approach for reducing specification assumptions in survey weighting. *arXiv preprint arXiv:2107.08075*, 2021.

J. Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1): 153–161, 1979.

M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. A Structural Approach to Selection Bias. *Epidemiology*, 15(5):615–625, 2004.

D. Horvitz and D. Thompson. Sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952a.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260): 663–685, 1952b.

C. Infante-Rivard and A. Cusson. Reflection on modern methods: Selection bias-a review of recent developments. *International Journal of Epidemiology*, 47(5):1714–1722, 2018. ISSN 14643685. doi: 10.1093/ije/dyy138.

Institute for Health Metrics and Evaluation. Covid-19 vaccine hesitancy, 2021. Oct 3, 2021 https://vaccine-hesitancy.healthdata.org.

M. Isakov and S. Kuriwaki. Towards principled unskewing: Viewing 2020 election polls through a corrective lens from 2016. *Harvard Data Science Review*, 2(4), 2020.

S. Jackman, S. Ratcliff, and L. Mansillo. Small Area Estimates of Public Opinion: Model-Assisted Post-stratification of Data from Voter Advice Applications. 2019.

C. Jackson, M. Newall, and J. Yi. Axios Ipsos Coronavirus Index, 2021. URL https://www.ipsos.com/en-us/news-polls/axios-ipsos-coronavirus-index.

N. Jackson. How Brexit Polls Missed the "Leave" Victory, 2016. URL https://www.huffpost.com/entry/brexit-polls-missed_n_576cb63fe4b017b379f58610.

S. Keeter. Growing and Improving Pew Research Center's American Trends Panel. *Pew Research Center*, 2019. URL https://www.pewresearch.org/methods/2019/02/27/growing-and-improving-pew-research-centers-american-trends-panel/.

C. Kennedy. Key things to know about election polling in the United States. *Pew Research Center*, 2020. URL https://www.pewresearch.org/fact-tank/2020/08/05/key-things-to-know-about-election-polling-in-the-united-states/.

C. Kennedy and H. Hartig. Response rates in telephone surveys have resumed their decline. *Pew Research Center*, 2019. URL https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/.

C. Kennedy, A. Mercer, S. Keeter, N. Hatley, K. McGeeney, and A. Gimenez. Evaluating Online Nonprobability Surveys. *Pew Research Center*, 2016.

C. Kennedy, M. Blumenthal, S. Clement, J. D. Clinton, C. Durand, C. Franklin, K. McGeeney, L. Miringoff, K. Olson, D. Rivers, et al. An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, 82(1):1–33, 2018.

C. Kennedy, D. Popky, and S. Keeter. How Public Polling Has Changed in the 21st Century. *Pew Research Center*, 2023. URL https://www.pewresearch.org/methods/2023/04/19/how-public-polling-has-changed-in-the-21st-century.

C. Kern, Y. Li, and L. Wang. Boosted Kernel Weighting – Using Statistical Learning to Improve Inference From Nonprobability Samples. *Journal of Survey Statistics and Methodology*, pages 1–26, 2020.

L. Kish. *Survey Sampling*. Wiley, 1965. ISBN 0-471-10949-5.

L. Kish. Weighting for unequal Pi. *Journal of Official Statistics*, 8(2):183–200, 1992.

F. Kreuter *et al.* Partnering with Facebook on a university-based rapid turn-around global survey. *Survey Research Methods*, 14(2):159–163, 2020.

W. Kruskal and F. Mosteller. Representative sampling, I: Non-scientific literature. *International Statistical Review/Revue Internationale de Statistique*, pages 13–24, 1979a.

W. Kruskal and F. Mosteller. Representative sampling, III: The current statistical literature. *International Statistical Review/Revue Internationale de Statistique*, pages 245–265, 1979b.

W. Kruskal and F. Mosteller. Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review/Revue Internationale de Statistique*, pages 169–195, 1980.

S. Kuriwaki, S. Ansolabehere, A. Dagonel, and S. Yamauchi. The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level. *American Political Science Review*, page 1–18, 2023.

B. E. Lauderdale, D. Bailey, J. Blumenau, and D. Rivers. Model-based pre-election polling for national and sub-national outcomes in the US and UK. *International Journal of Forecasting*, 36(2):399–413, 2020.

D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

L. Leemann and F. Wasserfallen. Extending the use and prediction precision of subnational public opinion estimation. *American Journal of Political Science*, 61(4): 1003–1022, 2017.

K. Z. LeWinn, M. A. Sheridan, K. M. Keyes, A. Hamilton, and K. A. McLaughlin. Sample composition alters associations between age and brain structure. *Nature Communications*, 8(1), 2017.

M. Link. AAPOR Response to New York Times / CBS News poll, 2014. URL `https://aapor.org/statements/aapor-response-to-new-york-times-cbs-news-poll/`.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

R. J. Little, B. T. West, P. S. Boonstra, and J. Hu. Measures of the degree of departure from ignorable sample selection. *Journal of survey statistics and methodology*, 8(5): 932–964, 2020.

T. J. Littlejohns, J. Holliday, L. M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, J. D. Bell, C. Boultwood, R. Collins, M. C. Conroy, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications*, 11(1):2624, 2020.

S. Lohr. *Sampling: Design and Analysis*. Duxbury Press, 2nd edition, 2010. ISBN 978-0495105275.

H. Lu and A. Gelman. A Method for Estimating Design-based Sampling Variances for Surveys with Weighting, Poststratification, and Raking. *Journal of Official Statistics*, 19(2):133, 2003.

V. Mayer-Schönberger and K. Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

X.-L. Meng. A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). *Past, Present, and Future of Statistical Science*, pages 537–562, 2014.

X.-L. Meng. Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations Big Data Paradox, and the 2016 US Presidential Election. *The Annals of Applied Statistics*, 12(2):685–726, 2018.

X.-L. Meng and X. Xie. I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econometric Reviews*, 33(1-4):218–250, 2014.

A. Mercer, A. Lau, and C. Kennedy. For Weighting Online Opt-In Samples, What Matters Most? *Pew Research Center*, 2018.

K. L. Miller et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19(11):1523–1536, 2016.

N. Mock, G. Singhal, W. Olander, J.-B. Pasquier, and N. Morrow. mVAM: A new contribution to the information ecology of humanitarian work. *Procedia engineering*, 159:217–221, 2016.

S. L. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)*. Cambridge University Press, second edition, 2007.

M. R. Munafò, K. Tilling, A. E. Taylor, D. M. Evans, and G. D. Smith. Collider scope: When selection bias can substantially influence observed associations. *International Journal of Epidemiology*, 47(1):226–235, 2018.

B. Murthy et al. Disparities in COVID-19 vaccination coverage between urban and rural counties: United States, December 14, 2020 – April 10, 2021. *Morbidity and Mortality Weekly Report*, 2021.

NatCen Social Research. Health Survey for England, 2018. URL `http://doi.org/10.5255/UKDA-SN-8334-1`. University College London, Department of Epidemiology and Public Health. UK Data Service.

K. H. Nguyen, P.-J. Lu, S. Meador, M.-C. Hung, K. Kahn, J. Hoehner, H. Razzaghi, C. Black, and J. A. Singleton. Comparison of COVID-19 vaccination coverage estimates from the Household Pulse Survey, Omnibus Panel Surveys, and COVID-19 vaccine administration data, United States Centers for Disease Control and Prevention, March 2021. 2021. `https://www.cdc.gov/vaccines/imz-managers/coverage/adultvaxview/pubs-resources/covid19-coverage-estimates-comparison.html`.

T. Q. Nguyen, A. Dafoe, and E. L. Ogburn. The magnitude and direction of collider bias for binary variables. *Epidemiologic Methods*, 8(1):20170013, 2019.

E. A. Nohr and Z. Liew. How to investigate and adjust for selection bias in cohort studies. *Acta Obstetricia et Gynecologica Scandinavica*, 97:407–416, 18.

D. K. Park, A. Gelman, and J. Bafumi. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4): 375–385, 2004.

D. K. Park, A. Gelman, and J. Bafumi. State-level opinions from national surveys: Poststratification using multilevel logistic regression. *Public Opinion in State Politics*, pages 209–28, 2006.

J. Park, C. Kim, and S. Son. Disparities in food insecurity during the COVID-19 pandemic: A two-year analysis. *Cities*, 131:104003, 2022.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995a.

J. Pearl. From Bayesian networks to causal networks. In *Mathematical models for handling partial knowledge in artificial intelligence*, pages 157–182. Springer, 1995b.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

D. Pennay, M. Goot, P. Hughes, D. Neiger, J. Sheppard, and J. Stirton. Discussion paper: disclosure standards for election and political polling in Australia. 2020a.

D. Pennay, M. Goot, D. Neiger, D. Trewin, P. J. Lavrakas, J. Stirton, P. Hughes, J. Sheppard, and I. L. McAllister. Inquiry into the performance of the opinion polls at the 2019 Australian Federal Election. 2020b.

Pew Research Center. The American Trends Panel survey methodology, 2020. URL `https://www.pewresearch.org/methods/u-s-survey-research/american-trends-panel/`.

B. Rader, C. M. Astley, K. Sewalk, P. L. Delamater, K. Cordiano, L. Wronski, J. M. Rivera, K. Hallberg, M. F. Pera, J. Cantor, et al. Spatial modeling of vaccine deserts as barriers to controlling sars-cov-2. *Communications Medicine*, 2(1):141, 2022.

A. Reinhart, E. Kim, A. Garcia, and S. LaRocca. Using the COVID-19 Symptom Survey to track vaccination uptake and sentiment in the United States, 2021. Sept 30, 2021 `https://perma.cc/GB72-C6Q5`.

D. Rivers. How the YouGov model for the 2017 General Election works, 2017. URL `https://yougov.co.uk/topics/politics/articles-reports/2017/05/31/how-yougov-model-2017-general-election-works`.

E. T. R. Rosenman, C. McCartan, and S. Olivella. Recalibration of Predicted Probabilities Using the "Logit Shift": Why Does It Work, and When Can It Be Expected to Work Well? *Political Analysis*, 31(4):651–661, 2023.

K. J. Rothman, J. E. Gallacher, and E. E. Hatch. Why representativeness should be avoided. *International Journal of Epidemiology*, 42(4):1012–1014, 2013.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

J. Rutenberg, K. Bensinger, and S. Eder. The 'Red Wave' Washout: How Skewed Polls Fed a False Election Narrative. *The New York Times*, 2022. URL `https://www.nytimes.com/2022/12/31/us/politics/polling-election-2022-red-wave.html`.

C. Ryan. Computer and Internet Use in the United States: 2016. *American Community Survey Reports*, ACS-39, 2017.

L. Saad. Historically Low Faith in U.S. Institutions Continues. *Gallup*, 2023. URL `https://news.gallup.com/poll/508169/historically-low-faith-institutions-continues.aspx`.

J. A. Salomon, A. Reinhart, A. Bilinski, E. J. Chua, W. La Motte-Kerr, M. M. Rönn, M. B. Reitsma, K. A. Morris, S. LaRocca, T. H. Farag, et al. The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proceedings of the National Academy of Sciences*, 118(51), 2021.

T. A. Santibanez, J. A. Singleton, C. L. Black, K. Nguyen, M.-C. Hung, S. Masalovich, P.-J. Lu, K. A. Brookmeyer, N. Abad, K. E. Barbour, A. Whiteman, B. P. Murthy, A. Wang, and H. A. Hill. Sociodemographic Factors Associated with Receipt of COVID-19 Vaccination and Intent to Definitely Get Vaccinated, Adults aged 18 Years or Above — Household Pulse Survey, United States, April 28–May 10, 2021. 2021. `https://perma.cc/C6KN-UDUY`.

C.-E. Särndal. Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 24(2):167, 2008.

B. Schaffner, S. Ansolabehere, and S. Luks. CCES Guide 2018.pdf. In *CCES Common Content, 2018*. Harvard Dataverse, 2019.

B. Settles. *Active Learning Literature Survey*. University of Wisconsin-Madison Department of Computer Sciences, 2009.

H. Shirani-Mehr, D. Rothschild, S. Goel, and A. Gelman. Disentangling Bias and Variance in Election Polls. *Journal of the American Statistical Association*, 113(522): 607–614, 2018.

N. Silver. Why FiveThirtyEight Gave Trump A Better Chance Than Almost Anyone Else. *FiveThirtyEight*, 2016. URL `https://fivethirtyeight.com/features/why-fivethirtyeight-gave-trump-a-better-chance-than-almost-anyone-else/`.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, second edition, 2001.

C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12 (3), 2015.

The Delphi Group. Topline Report on COVID-19 Vaccination in the United States, 2021. Carnegie Mellon University, in partnership with Facebook. `https://perma.cc/FKU8-LSYD`.

A. Tiu, Z. Susswein, A. Merritt, and S. Bansal. Characterizing the spatiotemporal heterogeneity of the COVID-19 vaccination landscape. *American Journal of Epidemiology*, 191(10):1792–1802, 2022.

X. M. Tu, X.-L. Meng, and M. Pagano. The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association*, 88(421):26–36, 1993.

R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In *NeurIPS 2020 Competition and Demonstration Track*, pages 3–26. PMLR, 2021.

US Census Bureau. Methodology for the United States population estimates: Vintage 2019, 2019. URL `https://perma.cc/PCC4-V48Q`.

US Census Bureau. Source of the Data and Accuracy of the Estimates for the Household Pulse Survey – Phase 3, 2021. URL `https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/Phase3_Source_and_Accuracy_Week_27.pdf`.

US Centers for Disease Control and Prevention. Trends in number of COVID-19 vaccinations, 2021a. URL `https://covid.cdc.gov/covid-data-tracker/#vaccination-trends`.

US Centers for Disease Control and Prevention. Estimates of vaccine hesitancy for COVID-19, 2021b. Sept 30, 2021 `https://data.cdc.gov/stories/s/Vaccine-Hesitancy-for-COVID-19/cnd2-a6zw/`.

US Centers for Disease Control and Prevention. COVID-19 Vaccination Trends in the United States, National and Jurisdictional, 2021c. URL `https://perma.cc/B5GH-C5UW`.

US Centers for Disease Control and Prevention. Reporting COVID-19 vaccination demographic data, 2021d. URL `https://perma.cc/H8A5-D7RX`.

R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. Walker, H. Fu, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, 20(6):669–677, 2020.

J. R. Wagner. *Adaptive survey design to reduce nonresponse bias*. PhD thesis, University of Michigan, 2008.

W. Wang, D. Rothschild, S. Goel, and A. Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.

Y. Wang, J. B. Holt, X. Zhang, H. Lu, S. N. Shah, D. P. Dooley, K. A. Matthews, and J. B. Croft. Comparison of Methods for Estimating Prevalence of Chronic Diseases and Health Behaviors for Small Geographic Areas: Boston Validation Study, 2013. *Preventing Chronic Disease*, 14:1–10, 2017.

D. Wasserman, S. Andrews, L. Saenger, L. Cohen, A. Flinn, and G. Tatarsky. 2020 National Popular Vote Tracker, 2020. URL `https://cookpolitical.com/2020-national-popular-vote-tracker`.

B. T. West, J. Wagner, S. Coffey, and M. R. Elliott. Deriving Priors for Bayesian Prediction of Daily Response Propensity in Responsive Survey Design: Historical Data Analysis Versus Literature Review. *Journal of Survey Statistics and Methodology*, 11 (2):367–392, 2023.

A. Wiśniowski, J. W. Sakshaug, D. A. Perez Ruiz, and A. G. Blom. Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8(1):120–147, 2020.

S. Yang, J. K. Kim, and R. Song. Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):445–465, 2020.

YouGov. About Our Panel, 2021. URL `https://today.yougov.com/about/about-the-yougov-panel/`.

X. Zhang, J. B. Holt, H. Lu, A. G. Wheaton, E. S. Ford, K. J. Greenlund, and J. B. Croft. Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *American Journal of Epidemiology*, 179(8):1025–1033, 2014.