A Framework for Estimating Individual Causal Effects

Patrick Lam

September 9, 2013

1 Introduction

What is the effect of political institutions on economic growth? Does UN intervention shorten the length of wars? Do job training programs increase wages and employment prospects? Does aspirin lower blood pressure? Researchers and scholars in every facet of industry and science grapple with causal questions all the time, using randomized studies and/or observational data to answer these questions. Almost always, the answers come in the following form: "there is a positive/negative causal effect[1] of the treatment on the outcome *on average*." Almost all research focuses on estimating the average causal effect, which is defined as the average of all the causal effects for every individual.[2] Yet in almost all cases, the average causal effect is not a specific causal effect for any one individual. Thus, there is a strong disconnect between what researchers generally measure (the average causal effect) and their actual quantity of interest (the causal effect for person i or country j).

The causal literature is quite clear on the difference between the average and individual-level causal effects. Under the potential outcomes framework, which dates back to Neyman but was formally defined and popularized under the "Rubin Causal Model" (Rubin, 1974), let W be a binary treatment variable taking on a value of 1 if a unit receives treatment and 0 if it receives control. The potential outcomes $Y(1)$ and $Y(0)$ represent the unit's outcome if it had received either treatment or control. The **individual causal effect (ICE)** for individual i is simply the difference between its potential outcomes under treatment and control.

$$\tau_i = Y_i(1) - Y_i(0)$$

Since at most one of the potential outcomes for each unit is observed, one cannot observe the causal effect of the treatment on the outcome. Rubin (1978) and Holland (1986) refer to this as the *fundamental*

[1] I use "treatment effect" and "causal effect" interchangeably throughout.
[2] I use the terms individual, observation, and unit interchangeably throughout.

problem of causal inference.

Almost every causal inference introduction begins with the ICE, yet quickly moves on to ways of identifying the average treatment effect (ATE).

$$\begin{aligned} \tau_{ATE} &= E[Y(1) - Y(0)] \\ &= E[Y(1)] - E[Y(0)] \end{aligned}$$

The ATE is easier to identify because one only needs to identify the means of the marginal distributions of the two potential outcomes. Standard regression techniques that are widely used and easy to implement have made the ATE the default quantity of interest. However, I argue that the focus on the ATE and various other average effects, while easier to estimate, loses a lot of potential information about treatment effect heterogeneity and has important implications for both research and policy. In this paper, I present a unified framework for estimating individual causal effects using many of the same tools already in place for estimating ATEs. I argue for a reorientation of the causal inference literature back toward estimating individual causal effects and expound upon the benefits of such an approach.

2 The Case for ICEs

Consider the following two statements:

- The treatment effect of W on Y is $\hat{\tau}$.
- Our model predicts that an increase of one unit of W increases Y by $\hat{\beta}$.

Variations of both statements are standard ways of describing causal effects in studies where a treatment variable W purportedly affects an outcome of interest Y. Whether the treatment effect is estimated from a regression model, from an experimental design, or from other forms of estimation, the estimate is usually some average treatment effect, yet the language is often unclear as to the units of interest. In the two statements above, $\hat{\tau}$ and $\hat{\beta}$ are average treatment effects, but it is important to note what average treatment effects represent. An ATE is not the effect of treatment on any one individual in the data (in most cases). An ATE is not the effect of treatment on a hypothetical individual with a given set of covariates. An ATE is not the effect of treatment for an average individual. Strictly speaking, an ATE is simply the average of all the individual effects for the individuals in the data. By reframing ATEs and

other causal quantities in terms of aggregations of individual effects, estimating ICEs can ***allow for a more precise and clear understanding of causal inference***. Possible confusion over what ATEs represent can be cleared up by referring to them as average effects of certain groups of individuals.

Often times, there is a temptation to apply the ATE to individuals of interest, such as in the case of using a regression coefficient to predict outcomes for future or counterfactual observations. There is a disconnect between what researchers are interested in, which is the effect for certain individuals or groups of individuals, and what researchers estimate, which is an average effect. For example, academics may be interested in explaining the effect of treatment in certain individuals, while policymakers may be interested in predicting the treatment effect for certain individuals. Rarely are researchers actually interested in "the average effect" per se. Estimating ICEs can ***reconcile the difference between what researchers estimate and what they are interested in***. Average effects only apply to individuals if researchers make the assumption of a constant treatment effect across individuals, which is a strong and usually unrealistic assumption. This leads to another point of emphasis between ICE estimation and ATE estimation, which is the ability of the former to examine treatment effect heterogeneity.

Consider the following study in Table 1 of a binary treatment indicator W on outcome Y with six observations. In Table 1a, the data are presented in a traditional setup where Y denotes the observed outcomes. In Table 1b, the same data are now presented in the form of potential outcomes. The

Table 1: A Study with Six Observations

i	W_i	Y_i
1	1	15
2	0	10
3	0	15
4	1	8
5	1	10
6	0	8

(a) Data

i	W_i	$Y_i(1)$	$Y_i(0)$
1	1	15	?
2	0	?	10
3	0	?	15
4	1	8	?
5	1	10	?
6	0	?	8

(b) Data with Potential Outcomes

question marks represent unobserved data, so one can think about causal inference as simply a missing data problem where the missing data are the unobserved potential outcomes for each unit i. A standard causal inference study would proceed to estimate the ATE with mild assumptions simply as

$$\begin{aligned} \hat{\tau}_{ATE} &= \bar{Y}_t - \bar{Y}_c \\ &= 11 - 11 \\ &= 0 \end{aligned}$$

where $\bar{Y}_t$ and $\bar{Y}_c$ denote the average observed outcomes for individuals receiving treatment and control respectively. The researcher would then note that the treatment has no effect. In a completely randomized experiment, this estimate is an unbiased estimate of the ATE since it is assumed that the observed potential outcomes are a random sample from the marginal distributions of the potential outcomes.

Now consider the same study in two different hypothesized worlds depicted in Table 2. In both scenarios, the missing potential outcomes are filled in (*italicized*) by drawing from the observed potential outcomes. The ATE remains the same as above in both cases. The last column of both tables contain the ICEs (**bolded**). If the researcher proceeded by estimating the ATE, the estimate would be unbiased

Table 2: Two Different Scenarios with Identical Average Treatment Effects

i	W_i	$Y_i(1)$	$Y_i(0)$	τ_i
1	1	15	*15*	**0**
2	0	*10*	10	**0**
3	0	*15*	15	**0**
4	1	8	*8*	**0**
5	1	10	*10*	**0**
6	0	*8*	8	**0**

(a) Treatment Has No Effect for Everybody

i	W_i	$Y_i(1)$	$Y_i(0)$	τ_i
1	1	15	*10*	**5**
2	0	*15*	10	**5**
3	0	*8*	15	**-7**
4	1	8	*15*	**-7**
5	1	10	*8*	**2**
6	0	*10*	8	**2**

(b) Treatment Helps Some and Hurts Some

and equal to 0 in both cases. However, the two scenarios are dramatically different. In Table 2a, the treatment has no effect for every individual. In Table 2b, the treatment has a large positive effect for some and a large negative effect for others. One may be tempted to conclude that an ATE of 0 implies the first scenario, but the second scenario is just as likely. With any given ATE value, there are an infinite number of ways in which the ICEs can aggregate to the same ATE. When estimating an ATE, researchers cannot say anything about effects for specific individuals or groups of individuals without further assumptions. Often, researchers use language that implies a constant effect for all individuals when only the ATE is estimated. In the presence of treatment effect heterogeneity, the ATE is a misleading quantity that hides much of what goes on in the data. By looking directly at ICEs, researchers can ***explore and discover treatment effect heterogeneity*** in a straightforward manner and explore any potential outliers or different underlying causal mechanisms amongst individuals or groups of individuals. The heterogeneity of treatment effects across individuals has important implications for research and policy-making.

Estimating ICEs also allows researchers to ***bridge the divide between quantitative and qualitative studies*** that exists in many areas of social science. As King, Keohane and Verba (1994) note, "the same logic of inference underlies both good quantitative and good qualitative research designs," yet there is still a disconnect between quantitative and qualitative scholars over which type of study is

better and which results are more reliable. Part of the disconnect exists because quantitative studies use large N statistical analyses to estimate causal effects whereas qualitative studies focus more on causal mechanisms and look closely at a small number of cases. I argue that another part of the disconnect stems from the different estimands and claims that each type of study attempts to make. Quantitative studies tend to collect data for a large N population, estimate average effects, and then implicitly attempt to apply the average effects to explain individual cases. Qualitative studies collect data for a small n sample, estimate individual or small n average effects, and implicitly attempt to generalize to the entire population. Each side estimates a different estimand, yet both attempt to address general average and specific individual effects. The results can often be dissatisfying to both sides, which leads to a divide. By estimating ICEs, quantitative researchers can speak directly to qualitative researchers about treatment effects on individual cases without sacrificing the ability to estimate average effects.

Although the inability to observe individual causal effects is the "fundamental problem of causal inference", one point that is seldom addressed is that the ICEs are fundamental to causal inference. If one can observe or estimate the ICEs, then any other causal estimand can be observed or estimated with very little effort. Thus, an additional benefit of focusing on estimating ICEs is that ***once the ICEs are estimated, the researcher can estimate any other causal effect by simply aggregating the ICEs***. Typically, if the researcher wants to estimate multiple causal estimands, he would have to develop a new model for each. By focusing on estimating the fundamental quantity in causal inference, researchers are able to estimate an unlimited number of other estimands by simple aggregation.

I have argued that there are at least five benefits to focusing on estimating ICEs rather than ATEs.

1. ICEs allow for a more precise and clear understanding of causal inference
2. ICEs reconcile the difference between the quantity in which the researcher is interested and the quantity the researcher estimates
3. ICEs allow researchers to explore and discover treatment effect heterogeneity
4. ICEs bridge the quantitative-qualitative divide
5. ICEs allow for easy estimation of every other causal estimand

Estimating ICEs, however, entails a cost because they are unidentified and much harder to estimate correctly. I argue that one can borrow existing techniques and frameworks in the causal inference and missing data literature to tackle the problem of estimation.

3 Existing Approaches to Causal Inference

Consider the typical situation in data analysis where there is a sample of N units indexed by i sampled from a large or infinite population.[3] Each unit i receives treatment W_i, where $W_i = 1$ indicates i received treatment and $W_i = 0$ indicates i received control. Each unit also has potential outcomes $Y_i(1)$ and $Y_i(0)$, where Y_i is the observed potential outcome depending on the value of W_i. Each unit also has a set of pretreatment covariates X_i which are assumed to be exogenous. Two basic assumptions are often needed to estimate causal effects:

Assumption 1: Ignorability of Treatment Assignment

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

This assumption is satisfied with random assignment of treatment or when X contains all pretreatment confounders that affect both W and the potential outcomes $Y(1), Y(0)$. Along with the ignorability of treatment assignment usually comes an assumption that $0 \leq P(W|X) \leq 1$, namely that there is positive probability of treatment for any X. The second important assumption is SUTVA.

Assumption 2: Stable Unit Treatment Value Assumption (SUTVA)

1. treatment assignment for one unit does not affect the potential outcomes of another (no interference or spillover effect):

$$(Y_i(1), Y_i(0)) \perp W_j, \quad \forall i \neq j$$

2. only one version of each treatment possible for each unit

With this basic setup, I now review the causal inference literature and different approaches used to estimate different causal estimands.

[3] The causal inference literature often uses the words sample, population, and superpopulation in different applications. Generally speaking, the sample is drawn from a population of a given size. Sometimes, the population is the sample, in which case the population is drawn from a larger superpopulation. For simplicity, I will generally refer to the data as the sample drawn from a very large or infinite population, but one can also think of the framework as a sample drawn from a superpopulation if the size of the sample is very close or equal to the size of the population.

3.1 Average Treatment Effects

Imbens (2004) provides an in-depth review of the literature of estimating average treatment effects, which I briefly review here. The most basic average treatment effect (ATE) that researchers estimate is simply

$$\tau^p_{ATE} = E[Y(1) - Y(0)]$$

The expectation here is over the population that the sample was drawn from. The more accurate definition for this estimand is the population average treatment effect (PATE), which differs from the sample treatment effect (SATE).

$$\tau^s_{ATE} = \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]$$

Since all the information known about PATE is captured in SATE, an estimator for SATE is the best and often a good estimator for PATE. Assuming that the sample is a random or representative sample from the population, the difference between SATE and PATE is in the variance of the estimates. Even if all the potential outcomes for the sample were observed, the potential outcomes for units not in the sample are not observed, so the variance needs to be adjusted upward for PATE. In most cases, researchers are interested in the population estimands, although the sample estimands can be of interest in situations where the sample is not representative of the population. In reviewing the causal inference literature, I ignore the differences between the sample and population versions of the estimands, assuming that researchers are estimating sample estimands with possible adjustments to estimate population estimands.

If treatment assignment is randomized or plausibly randomized such as in an experiment, then researchers can estimate the ATE by a simple difference in means,

$$\hat{\tau}_{ATE} = \bar{Y}_t - \bar{Y}_c$$

where Y_t and Y_c denote outcomes for observations that received treatment and control respectively.

3.1.1 Regression Approaches

Short of treatment assignment randomization, researchers need to condition on the set of confounders X to estimate the ATE. Perhaps the most common class of methods to condition on X is the class of regression estimators, which uses some functional form to estimate the average potential outcomes

$\mu_{(w)}(x)$ given $X = x$ for $w = 1, 0$. The general form of the regression estimator averages over the empirical distributions of the covariates for treatment and control groups:

$$\hat{\tau}_{ATE,reg} = \frac{1}{N}\sum_{i=1}^{N}[\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)]$$

Many regression estimators impose a functional form for $\hat{\mu}_w(X_i)$ and possibly a parametric distribution for Y. The common linear model imposes a linear relationship between X and $\mu_w(X)$

$$\begin{aligned} \mu_{(w)}(X_i) &= \alpha + \tau W_i + \beta' X_i \\ Y_i &= \alpha + \tau W_i + \beta' X_i + \epsilon_i \end{aligned}$$

and estimates the parameters by ordinary least squares. Generalized linear models (McCullagh and Nelder, 1989) specify the relationship between X and $\mu_{(w)}(x)$ through a linear functional form and a link function $g(\cdot)$ and also impose a parametric distribution $f(\cdot)$ on Y.

$$\begin{aligned} g(\mu_{(w)}(X_i)) &= \alpha + \tau W_i + \beta' X_i \\ Y_i &\sim f(\cdot|\mu_{(w)}(X_i)) \end{aligned}$$

Other regression models, such as kernel regression, generalized additive models, smoothing splines, local polynomial regression, are semiparametric or nonparametric and relax the parametric and linearity assumptions. The literatures on these models is enormous and I will not review it here (see Hastie, Tibshirani and Friedman (2009) for a extensive introduction). Although regression is commonly used to reduce bias and increase precision in estimating ATEs, it can actually lead to more bias when the functional form of the covariates is specified incorrectly. In the case where there is little covariate overlap between treatment and control groups, regression results can be very dependent on model specifications.

3.1.2 Matching Approaches

Another way to condition on X is to use matching methods, which first appeared in the early 20th century but was not developed theoretically until the 1970s (Rubin, 1973*a*,*b*). Unlike regression methods, matching methods rely less on functional form and model assumptions. The goal of matching is to approximate a randomized experiment by matching individuals from treatment and control groups with similar covariate profiles. Observations that do not have overlap in covariates are removed from the matched sample to avoid extrapolation. In the ideal matching scenario, each treatment observation

would be matched with one or more control observations with the same exact values on all the covariates and/or vice versa. The average treatment effect would then be calculated by differencing the treatment and control outcomes from this matched sample. This exact matching approach may be feasible in the case of a small number of discrete covariates. However, if there are continuous covariates and/or as the number of covariates increases, exact matching is not feasible in a finite sample because of the curse of dimensionality. Numerous matching methods have been developed to match observations in the hopes of achieving covariate balance across treatment and control groups. Stuart (2010) provides a comprehensive overview of the current matching methods developed. One point to note is that when estimating average treatment effects, the only requirement is that the distributions of the covariates for the treated and control groups be similar in the matched sample, which is less restrictive than requiring close or exact matches on all the variables for all observations. Researchers can then combine the matched sample with regression analysis to adjust for remaining imbalance after matching. Using the two methods in combination also helps to form "doubly robust" estimators which are less sensitive to misspecifications in either the matching method or regression model (Rubin, 1973*b*, 1979; Ho et al., 2007).

Researchers who use matching methods to estimate average treatment effects do so by matching each treatment observation to one or more control observations and each control observation to one or more treatment observations. Often researchers are interested in another causal estimand, the average treatment effect for the treated (ATT):

$$\tau_{ATT} = \frac{1}{N_t} \sum_{i:W_i=1} [Y_i(1) - Y_i(0)]$$

where $N_t = \sum_{i=1}^{N} W_i$ is the number of treated units. From a computational standpoint, estimating the ATT is simpler since the researcher only needs to match the treated units with control units and does not have to match the control units with treated units or worry about whether the best matches for one imply best matches for the other.. From a policy and academic standpoint, since the treatment effect is of interest, it may be more appropriate to only look at units that actually received the treatment. Depending on the nature of treatment assignment, the treated group may be qualitatively different than the control group and thus important to look at separately. Although less common, the average treatment effect for the controls (ATC) may also be of interest:

$$\tau_{ATC} = \frac{1}{N_c} \sum_{i:W_i=0} [Y_i(1) - Y_i(0)]$$

where $N_c = \sum_{i=1}^{N} (1 - W_i)$ is the number of control units. If ATT = ATC, then the ATE = ATT = ATC.

Also note that when using matching, observations which do not have good matches may be discarded, which changes the quantity of interest being estimated.

When implementing a matching method, the researcher has to make several choices. At each step, there are many options that the researcher can choose from. The factors to consider in any matching method are:

1. **The variables to include in the matching**

   Since matching is a way of conditioning on confounding variables to satisfy ignorability of treatment assignment, the researcher should include all pre-treatment variables that affect treatment assignment and the outcome. However, the curse of dimensionality almost certainly implies that covariate balance will be harder to achieve as the number of variables to match increases. With many variables to match on, improving balance on one variable may very well decrease balance on another and increase the bias of the estimate. Researchers may have to make choices on which variables to prioritize or choose matching methods that put different weights on different variables (Diamond and Sekhon, 2013). One type of variable that generally should not be included is any variable that is affected by the treatment. Including these post-treatment variables can result in bias in the estimate (Rosenbaum, 1984).

2. **The measure of closeness between observations**

   When balancing on multiple variables (especially in the presence of continuous variables), the curse of dimensionality makes it difficult to determine how "close" observations are on the covariates. Researchers need to determine a measure of distance between observations and also decide how to match observations given their distances. In the ideal case of exact matching, observations are matched if all their covariate values are the same. Exact matching is rarely feasible, but one strategy is to coarsen the covariates and match exactly on the coarsened variables (Iacus, King and Porro, 2012). Another strategy is to define distance between observations by a one-dimensional balancing score for each observation that summarizes the information in the covariates. Some examples of balancing scores include the Euclidean distance, the Mahalanobis distance, propensity scores (Rosenbaum and Rubin, 1983), and prognostic scores (Hansen, 2008).

Once distance is defined, researchers must then choose how to convert the distances into matches. One option is to do nearest-neighbor matching, where each treated observation is matched with its closest neighbors. The algorithm for nearest-neighbor matching may be greedy, with each observation choosing its matches in order, or optimal, taking into account all possible matches and minimizing a global distance measure. Note that greedy algorithms depend on the order of the observations. Another option is to divide the observations into a number of subclasses based on their distance measures, where each subclass contains at least one treatment and one control observation. Observations in the same subclass are then matched. Deciding the number and boundaries of the subclasses themselves is another choice for the researcher. The researchers can define these directly, indirectly as in the case of coarsened exact matching (Iacus, King and Porro, 2012), or through an algorithm as in the case of full matching (Rosenbaum, 1991). A third option is to match using the whole set of observations but weighting the observations by their distance measure as in Imbens (2000). Hainmueller (2012) uses entropy balancing to derive weights, optimizing balance on the sample moments of the covariate distributions. Note that all the options can be considered as weighted matching, where the first two options put weights of either 0 or 1 on every observation. One can also combine any of these options with calipers, which place restrictions on the distances for acceptable matches.

3. **The number of observations to serve as the "donor pool"**

For each observation, the researcher chooses to match it with one or more "donor" observations that received the opposite treatment. When matching each treated observation with control observations, all the control observations represent the donor population from which the researcher chooses M of them for the donor pool. The size of the donor pool in M-to-1 matching is often an arbitrary choice by the researcher. The most common choice is 1-to-1 matching where the closest observation on the distance measured is chosen. The choice of M is often a trade-off between bias and variance. In the case of an unlimited pool of exact matches, increasing M reduces the variance of the estimate by including more observations with more information. However, in practice, there are rarely exact matches, so increasing M results in matching on observations that are farther away on the distance measure. This decrease in variance by increasing M comes at the cost of increasing bias from matching on less similar observations (Rosenbaum and Rubin, 1985). M is also often chosen indirectly, such as in methods using strata or subclassification where M is determined by the number of donor observations in the subclass or in methods using weights where the number

of donor observations is determined by the weights. M can also be allowed to vary, which may reduce bias even further (Ming and Rosenbaum, 2000).

4. **Whether to match with or without replacement**

   When the number of possible donor pool observations is relatively small, researchers have an option to match with replacement. Matching with replacement reuses observations in multiple donor pools such that certain observations may be matched more than once. Matching with replacement can reduce bias since it usually results in better quality donor pools. However, the outcome analysis should take into account the fact that observations are used multiple times. The number of unique donor observations used should also be monitored so that the results are not dependent on using information from a small number of the donor population.

5. **How to check covariate balance to determine the success of the matching**

   The goal of matching is to create a matched dataset with similar distributions in the covariates for the treated and control groups. Therefore, to verify that the matching worked properly, the researcher must assess covariate balance in the matched sample such that $\tilde{p}(X|W=1) = \tilde{p}(X|W=0)$ where $\tilde{p}$ is the empirical distribution. Ideally one would like to examine the multivariate distributions of the covariates for the treatment and control groups. However, comparing multivariate distributions becomes difficult as the dimensions increase. Although some have suggested using multivariate imbalance measures (such as the $\mathcal{L}1$ statistic), most applications look at the marginal empirical distributions of the covariates and check balance on the moments (such as the standardized means) of the distributions. Others visualize balance graphically with Q-Q plots or plots of the different moments of the distributions. Other ways to check balance include running hypothesis tests to test whether the marginal distributions of the treated and control group are the same, although Imai, King and Stuart (2008) argue rightly that what matters is the in-sample balance rather than out-of-sample population balance. There are also certain matching methods that allow the researcher to define the level of imbalance ex-ante, thus constraining the post-matching imbalance to a certain level.

Matching methods have become increasingly popular in the causal inference literature because of its ability to mimic randomized experiments and its lesser reliance on parametric modeling assumptions. I

revisit many of these matching methods in more detail and discuss how matching methods can be used to estimate individual causal effects.

3.1.3 Other Approaches

Besides matching and regression, other approaches exist that try to identify average treatment effects, usually by leveraging aspects of the data or external circumstances to approximate random assignment of treatment. For example, one can use natural experiments where a treatment has been pseudo-randomized by nature. Another approach is to use instrumental variables analysis, where the researcher has a randomized or plausibly ignorable instrumental variable that is correlated with the treatment variable of interest. Finally, regression discontinuity designs attempt to leverage sharp discontinuities in the treatment variable to conduct analyses as if the treatment has been randomized for units near the discontinuity.

A lesser known but potentially powerful modeling approach to estimate average treatment effects is with Bayesian methods. Rubin (1978) introduces a general Bayesian framework for estimating treatment effects. One way to use Bayesian methods is to model the potential outcomes directly. Another way is to estimate regression models using priors on the regression coefficients to weight the importance of various covariates. Although very little has been done on integrating matching methods and Bayesian approaches, I argue for using a Bayesian framework with matching methods to estimate individual causal effects.

3.2 Treatment Effect Heterogeneity

Treatment effect heterogeneity exists when there are varying average treatment effects for various subgroups of the inferential population. Treatment effect heterogeneity is an important topic in many fields, especially in the medical sciences where a treatment may help some patients but hurt others (Kravitz, Duan and Braslow, 2004; Rothwell, 2005). Political scientists are also increasingly interested in treatment effect heterogeneity with substantive implications (Feller and Holmes, 2009; Arceneaux and Nickerson, 2009; Gaines and Kuklinski, 2011; Imai and Strauss, 2011). In the presence of treatment effect heterogeneity, estimating a simple average treatment effect may mask important differences in treatment effects as I demonstrated above. The most common way to test for treatment effect heterogeneity is to estimate the average treatment effect for different subgroups of the sample using any of the methods described

above. The subgroups are defined by the specific covariates and the average treatment effect within a subgroup is commonly known as the **conditional average treatment effect (CATE)**:

$$\tau_{CATE,x} = E[Y(1) - Y(0)|X = x]$$

where x denotes the covariate values of the subgroup. Treatment effect heterogeneity occurs when the CATEs differ for different subgroups. However, two general sets of complications arise when estimating multiple CATEs: 1) small sample sizes and limited power and 2) multiple testing problems and arbitrarily defined subgroups.

Recall that for any statistical test, the power of the test is inversely related to the sample size. When testing for effects within subgroups in the same dataset, the sample size is usually significantly smaller than the size of the original dataset N. This is especially true in clinical trials, where N is usually small to begin with. Unless the subgroup treatment effects are quite large, standard statistical tests often fail to detect effects in subgroups (Pocock et al., 2002). One solution to the problem of small sample sizes in subgroup analyses is to use interaction terms where the variable defining the subgroups is interacted with the treatment indicator. Although the use of interaction terms better captures the extent of the information in the data and uses the data more efficiently, the estimators used are still usually limited by the need to appeal to large sample properties, while the subgroup analyses rely on smaller and smaller samples.

Ironically, many existing subgroup analyses are also susceptible to a second complication of multiple testing problems and arbitrarily defined subgroups. When looking for treatment effect heterogeneity, the researcher often tests for significant effects over multiple subgroups defined by the covariates. With multiple tests, the probability of a false positive is greatly inflated and can lead to misleading results (Lagakos, 2006). Crump et al. (2008) develop nonparametric tests for the null of no treatment effect heterogeneity, which bypass the multiple testing problem but fail to specify exactly which subgroups have heterogeneity. In addition to the multiple testing problem, the choice of subgroups to examine for treatment effect heterogeneity is often left to the researcher, which creates potential validity and incentive compatibility concerns. Subgroups can be chosen either arbitrarily or with some substantive theory in mind. They can be prespecified before the experiment or chosen post-hoc. Recent data mining techniques have been developed to remove the choice of subgroups from the researcher's control by using learning algorithms to search through the space of treatment-covariate interactions to detect statistically significant effects (Green and Kern, 2012; Imai and Ratkovic, 2013).

The literature on subgroup analysis and treatment effect heterogeneity is relatively small compared to the literature on estimating ATEs. When testing for treatment effect heterogeneity, it is sometimes unclear whether the quantities of interest are the CATEs themselves or the differences in CATEs. Estimating CATEs often seems to boil down to estimating ATEs on smaller randomly chosen subsets of data. The estimators themselves often rely on large sample approximations that may not even hold in the larger full dataset. Matching techniques that often work well in estimating ATEs are seldom used in estimating CATEs. The interpretations of the interaction terms in the treatment effect heterogeneity setting may also be tricky, especially when the covariate that is interacted is more complicated than a binary variable. Other scholars have approached the topic differently by developing bounds for the proportion of the population that has treatment effect heterogeneity (Gadbury, Iyer and Albert, 2004). I argue that an even easier way to examine treatment effect heterogeneity is to estimate the individual causal effects themselves, bypassing the need for complicated interaction models and testing at the subgroup level.

3.3 Individual Causal Effects

The literature on estimating individual causal effects is substantially smaller than either the literature on estimating ATEs or treatment effect heterogeneity, which mirrors the lack of attention scholars have paid to the topic. The simplest way to estimate an ICE is to estimate a general model for ATEs and predict the individual effects based on that model. For example, in medicine, researchers suggest calculating the baseline disease risk for any individual patient based on covariates and an existing model and then calculate the effect of treatment on that patient using the overall effect from a clinical trial (Dorresteijn et al., 2011). A second approach to estimate ICEs requires multiple datapoints over time, usually one or more "pre-treatment" datapoints and one or more "post-treatment" datapoints. The simplest example would be a crossover design, where individuals are randomized to one treatment at time t and another at time $t + 1$. In this case, the individual would act as both treatment and control observations. However, strong assumptions about time-period effects and treatment carry-over effects across time need to be made. Steyer (2005) proposes a more general model involving multiple pre-treatment and post-treatment observations to measure the "latent" true expected outcomes. Abadie, Diamond and Hainmueller (2010) introduce the use of synthetic controls to estimate the treatment effect for a single unit with time-series data. The synthetic controls are created by comparing and weighting all the control units with the unit that received treatment and calibrating based on the outcome variables for the time periods before the unit received the treatment.

Recent work has focused on using both Bayesian methods and matching methods developed for estimating ATEs and adapting them to estimate ICEs. As Abadie and Imbens (2006) put it, any matching estimator simply "imputes the missing potential outcomes." Rubin and Waterman (2006) use propensity score matching to create "clones" for each treated unit in order to estimate ICEs, although their approach does not include any uncertainty estimates. An (2010) suggests that using a Bayesian propensity score estimator can incorporate uncertainty over the matching procedure to estimate individual effects. Rubin (2005) presents a general framework in which missing potential outcomes can be imputed by drawing from the posterior predictive distribution of potential outcomes in any Bayesian model. Pattanayak, Rubin and Zell (2012) stratify treatment and control observations using estimated propensity scores and then use a Bayesian model within each strata to estimate ICEs. Gutman and Rubin (2012) develop imputation methods using subclassification and splines with knots at the borders of the subclasses to impute the missing potential outcomes. Finally, Jin and Rubin (2008) assume that the potential outcomes $Y(1)$ and $Y(0)$ are correlated by the parameter ρ and test the sensitivity of the causal effects to different values of ρ. In the next section, I introduce a flexible general framework to estimating ICEs that builds on many of these studies, using both Bayesian methods and a wide variety of matching methods.

4 Estimating Individual Causal Effects

One reason why ICEs are not estimated or points of focus is that ICEs are not identified in the data without further assumptions. Suppose that for an individual i, one posits that the ICE can be -1000, 0, or 9999.8. Statistical identification requires that the data and our estimation method tell us which of the three values is more likely to be true. However, since one does not observe the missing potential outcome, the data cannot give us any more information about the ICE for individual i. Given that identification is impossible, I argue that one should estimate ICEs by deriving a range of plausible values for the ICEs given information from other observations in the data. I use a Bayesian framework which gives us a posterior distribution of our ICEs based on information from the data and our prior beliefs rather than an identified point estimate.

The approach I use builds on a Bayesian framework for imputing missing potential outcomes first introduced by Rubin (1978), with similarities to the approach used in Pattanayak, Rubin and Zell (2012). As before, let W_i denote a binary treatment assignment indicator for unit i with an observed outcome Y_i and a vector of pre-treatment covariates X_i. Define Y_i^{mis} to be the unobserved potential outcome for

unit i:

$$Y_i^{mis} = \begin{cases} Y_i(1) & \text{if } W_i = 0 \\ Y_i(0) & \text{if } W_i = 1 \end{cases}$$

Let τ_i be the individual causal effect for unit i:

$$\tau_i = \begin{cases} Y_i^{mis} - Y_i & \text{if } W_i = 0 \\ Y_i - Y_i^{mis} & \text{if } W_i = 1 \end{cases}$$

which I can rewrite simply as

$$\tau_i = W_i(Y_i - Y_i^{mis}) + (1 - W_i)(Y_i^{mis} - Y_i)$$

Since τ_i is a deterministic function of Y_i^{mis} and the observed data, I can calculate τ_i by simply imputing Y_i^{mis}. Our uncertainty around τ_i also comes only from our uncertainty around Y_i^{mis} since Y_i is observed.

To start, recall the most basic framework found in many regression models used in the social sciences (e.g. generalized linear models). In a typical regression setup, Y is a random variable that follows some probability distribution defined by a set of parameters θ conditional on covariates X and treatment W.

$$Y_i \quad \sim \quad f(\cdot|\theta_i, X_i, W_i)$$

The parameter vector θ_i includes the mean of Y_i, μ_i, which is usually parameterized as a function of the regression coefficients β, and possibly some ancillary parameters ϕ. I then estimate β in our regression model and derive average causal effects, since β is not subscripted by i. Note that typical regression models do not reference the missing potential outcomes, although one could use the regression model to predict the missing potential outcomes.

In my framework for estimating τ_i, I take a slightly different approach to modeling the data. Suppose instead that our data is a finite sample of size N drawn from the following data generating process:

$$\begin{aligned} Y_i &= h(X_i^{(p)}) \\ Y_i^{mis} &= h(X_i^{(p)}, \tau_i) \end{aligned} \quad \text{for } W_i = 0$$

$$\begin{aligned} Y_i &= h(X_i^{(p)}, \tau_i) \\ Y_i^{mis} &= h(X_i^{(p)}) \end{aligned} \quad \text{for } W_i = 1$$

where $X_i^{(p)}$ is the set of all prognostic variables (variables that predict the outcome) including any confounding variables and $h(\cdot)$ is some unknown function. First, note that the framework is restricted to the finite sample and one can only estimate individual causal effects for units in the data. Looking only at the finite sample allows us to appeal to a Bayesian setup. Also, the idea of individual causal effects is fundamentally restricted to the sample since individuals only appear in the data, and not in some superpopulation. I also assume that if the data generating process repeated multiple times under the same exact conditions, τ_i remains constant for i. Second, the potential outcomes are fixed and completely determined by $X_i^{(p)}$ and W, which are also fixed. In theory, if every single variable that affects the outcome can be measured, one could predict the outcome perfectly.[4] In practice, only a very small subset of $X_i^{(p)}$ is observed. Partition $X_i^{(p)}$ into a set of observed covariates, X_i, and a set of unobserved covariates, $X_i^{(u)}$.

$$X_i^{(p)} = \{X_i, X_i^{(u)}\}$$

If X_i contains at least all the variables that makes treatment assignment ignorable, then the ignorability assumption gives us

$$\begin{aligned} (Y(1), Y(0)) &\perp W|X \\ \tau &\perp W|X \\ X^{(u)} &\perp W|X \end{aligned}$$

Note that the assumption that τ is independent of treatment assignment conditional on X implies that one can use information from the opposite treatment group to inform the missing potential outcome for i. In a simple example, assume that observation i is treated and observation j is control and they have the same value for $Y(0)$. If, for example, the ICEs were systematically larger for those assigned control, then using information from j would overestimate τ_i.

These ignorability statements imply some type of randomness in the data. I assume that conditional on the observed X, the unobserved $X^{(u)}$ are essentially random across treatment and control observations.

[4]The approach I am taking to the data generating process is that any outcome can be predicted perfectly by observing the complete set of prognostic variables and knowing the functional form. Philosophically, this argument may conflict with the traditional statistical idea of randomness and unpredictability. In practice, the two approaches are the same since the full set of prognostic variables is never observed and I proceed by modeling the outcomes as random. However, I take this approach to make the two points. First, since the quantity of interest is the individual causal effect, I want to stress that the subscript i takes on a special meaning that is specific to that individual. Therefore, i can be modeled and predicted completely in theory. Second, I want to make the point that including more prognostic variables can give us more information about the missing potential outcome.

The randomness is then modeled with the following:

$$Y_i^{mis} \sim f(\cdot|\theta_i^{mis}, X_i, W_i)$$

where θ_i^{mis} represents the distributional mean of the outcomes conditional on the observed X_i.[5] Simply put, observations with the same values of X_i and W_i are randomly drawn from a common distribution, conditional on Assumptions 1 and 2 being satisfied. Strictly speaking, θ_i^{mis} should be denoted as θ_{X_i,W_i}^{mis}, which indicates that it is the mean of the missing potential outcome and that observations with the same observed covariate vector and treatment status as i have the same mean. I use θ_i^{mis} to simplify notation.

Consider an observation j where $X_j = X_i$ and $W_j = 1 - W_i$. Then this implies that Y_i^{mis} and Y_j are modeled as generated from the same distribution:

$$\begin{aligned} Y_i^{mis} &\sim f(\cdot|\theta_i^{mis}, X_i, W_i) \\ Y_j &\sim f(\cdot|\theta_i^{mis}, X_j = X_i, W_j = 1 - W_i) \end{aligned}$$

This suggests that if i is a treated observation, one can use observed outcomes for control observations with the same value on X as i to estimate θ_i^{mis}. This also implies that one can model the data generating process for the observed data as

$$\begin{aligned} Y_i &\sim f(\cdot|\theta_i^{obs}, X_i, W_i) \\ \theta_i^{obs} &= W_i(\theta_i^{mis} + \tau_i) + (1 - W_i)(\theta_i^{mis} - \tau_i) \end{aligned}$$

However, because I assume that Y_i is fixed and observed, θ_i^{obs} is not an interesting parameter and is not estimated. θ_i^{mis}, the mean of the distribution for the missing potential outcome, is the key parameter of interest in this framework. The stochastic nature of the outcomes reflects the contributions of the unmeasured prognostic variables, which are assumed to be independent of treatment assignment. In other words, each potential outcome for any individual i is a deterministic function of observed and unobserved prognostic covariates. Then θ_i^{mis} is estimated by matching to create observations that are considered to be similar on the observed covariates X:

$$\theta_i^{mis} = m(X_i, W_i, Y)$$

[5] For some distributions, there may be ancillary parameters in addition to the mean. In that case, θ_i^{mis} would be a vector of parameters. For the sake of notational convenience and simplicity, I assume that $f(\cdot)$ is parameterized solely by the mean for now.

where $m(\cdot)$ is a matching estimator. The assumption made with this setup is that the potential outcomes are independent conditional on X_i. That is, Y_i gives no extra information about Y_i^{mis} and vice versa.

There is a slight difference between my framework and other approaches to causal inference as to where the randomness occurs in the dataset. Most approaches make appeals to superpopulations and estimate population parameters. Units are assumed to be drawn from these superpopulations. For example, one common approach is to assume that W, X and Y are all random variables (Rubin, 2005, 2008). Abadie and Imbens (2006) on the other hand assume that the triplet $\{Y, W, X\}$ is drawn at random. In my approach, there is no superpopulation and the only randomness comes from the unknown $X^{(u)}$. I am strictly interested in estimands in the observed sample. If $X^{(p)}$ was fully observed, there would be no randomness and all the parameters can be calculated. Although my framework can be adjusted and applied to other approaches or appeal to superpopulations, I make explicit the notion that randomness in Y^{mis} comes only from not observing $X^{(u)}$. In practice, there is very little difference between my assumption about the source of randomness and the typical setup. For example, one can think of $X^{(u)}$ as simply the error term ϵ in linear regression models.

This framework involves two steps: a matching step to estimate θ_i^{mis} and an imputation step to get an imputed value of Y_i^{mis} accounting for the unobserved prognostic covariates. Each step is also characterized by a type of uncertainty that eventually propagates to uncertainty around τ_i. The matching step has *estimation uncertainty* and the imputation step has *fundamental uncertainty* (King, Tomz and Wittenberg, 2000). Estimation uncertainty refers to the uncertainty in estimating θ_i^{mis}, which encompasses uncertainty over the parameters of the matching procedure, uncertainty due to finite sample size, and possibly even uncertainty over the choice of the matching procedure itself. Estimation uncertainty is a function of the variation in outcomes and the size of the donor pool. Fundamental uncertainty is usually described as randomness or chance events that affect the outcome but is not included in the set of conditioning variables. In other words, fundamental uncertainty reflects the influence of our unmeasured prognostic variables. All things being equal, conditioning on more variables that affect the outcome should reduce fundamental uncertainty. I introduce various ways to perform the matching step in the next section, borrowing from many existing techniques in the causal inference literature. Both matching and imputation steps are then incorporated into a general Bayesian model. I then test the performance of the various techniques for estimating τ_i via simulation.

4.1 The Matching Step

To estimate τ_i, I first need to conduct matching N times to estimate θ_i^{mis} for all i in the data. Let $D_j^{(i)}$ be a binary variable that denotes whether or not an observation j is in the donor pool for observation i when estimating τ_i:[6]

$$D_j^{(i)} = \begin{cases} 1 & \text{if } W_j \neq W_i \ \& \ j \text{ is a match to } i \\ 0 & \text{otherwise.} \end{cases}$$

The size of the donor pool for observation i is simply $\sum_{j=1}^{N} D_j^{(i)}$. In matching procedures where observations can be weighted donors, $D_j^{(i)}$ acts as the donor weight and can take on any value between 0 and 1.[7] The matching step involves defining $D_j^{(i)}$ by choosing a set of donor observations that are similar to i on the conditioning variables X_i. I then use the observed outcomes in the donor pool to estimate θ_i^{mis}.

In an ideal world, one can expand X_i to include all prognostic covariates measured without error and the observations in the donor pool would be exact matches to i on all X_i. There would be no estimation or fundamental uncertainty and Y_i^{mis} can be imputed exactly. However, in practice, finite sample sizes, a large number of prognostic covariates, many of which are unobserved, and/or the presence of continuous covariates precludes the possibility of exact matching on all prognostic covariates. Instead, I use matching procedures to define $D_j^{(i)}$ and calculate the mean of the donor pool as

$$\bar{Y}_{D^{(i)}} = \frac{\sum_{j=1}^{N} Y_j D_j^{(i)}}{\sum_{j=1}^{N} D_j^{(i)}}$$

I then use a Bayesian model (described below) to combine $\bar{Y}_{D^{(i)}}$ and a prior to estimate θ_i^{mis}.

The decisions made with respect to the selection of the matching procedure mirrors the choices usually made when using matching to estimate average treatment effects. In this case, since the quantity of interest is the individual causal effect, the goal is no longer simply distributional balance across treatment and control observations. Instead, one needs to create a donor pool that is as close to i on X_i as possible.

The following choices must be made with respect to our matching algorithm:

[6] In the case of exact matching, $D_j^{(i)}$ denotes whether j and i are exact matches ($X_j = X_i$). Since most methods researchers use are approximate matching methods, $D_j^{(i)}$ is random even if W and X are fixed. One should conceptually think about $D_j^{(i)}$ as an indicator for whether or not $X_j \approx X_i$.

[7] For non-binary donor weights, some of the equations below must be adjusted. For now, I assume that $D_j^{(i)}$ only takes on a value of 0 or 1.

- **The set of conditioning variables** X: All confounding variables should be conditioned on to satisfy the ignorability of treatment assignment and causal effect independence assumptions. In addition, other prognostic variables should also be conditioned on to improve the efficiency of the estimates and possibly reduce bias (Rubin and Thomas, 2000; Pocock et al., 2002). However, with limited sample sizes and small donor pools, there is a tradeoff between finding good matches and conditioning on more variables, akin to a bias-variance tradeoff. Researchers should prioritize conditioning on confounders that are also highly predictive of the outcome.

- **Size of the donor pool**: The size of the donor pool, $M = \sum_{j=1}^{N} D_j^{(i)}$ is chosen either directly or indirectly by the researcher and may vary across i. By definition, increasing the size of the donor pool results in the inclusion of matches that are either worse or about the same in terms of similarity to i on X_i. This results in a more efficient estimate of θ_i^{mis}, but may also introduce more bias due to the inclusion of poorer quality matches.

- **Matching with or without replacement**: Since the quantity of interest is at the individual level, reusing matches for multiple ICEs does not pose any problems and leverages better information. Matching with replacement is ideal and may be necessary for small sample sizes.

- **Weighting donor observations**: By default, in most matching applications, donor observations each receive a weight of 1, implying that all donors are equally good matches. Expanding the size of the donor pool likely results in matches that are poorer matches, so the researcher can choose to downweight donors as a way to reduce the influence of poor matches on the estimate. This also reduces the effective size of the donor pool and incorporates greater uncertainty in the presence of poorer matches.

- **Definition of closeness**: Since in most cases, exact matching is impossible, choosing the definition of closeness between matches is probably the most important task. One can choose amongst a myriad of dimension-reducing balancing scores, although exact matching should be used when possible. A mix of exact matching and balancing scores is also feasible.

- **What to do with unmatched observations**: For some observations, it is likely that the pre-defined criteria produces no matches for the donor pool. For estimating individual causal effects, discarding unmatched observations means not estimating a causal effect for that individual. When aggregating to average effects, discarding observations changes the quantity of interest. The researcher can force matches by relaxing some of the matching criteria imposed.

Short of exact matching, it is unclear which matching procedure performs the best a priori in estimat-

ing θ_i^{mis}. I consider a few options that are prevalent in the matching literature, adapting and combining some of them to try to gain efficiency and reduce bias. I then test the performance of each of these options via simulation. My Bayesian model also includes an option to incorporate uncertainty around any parameters within a specific matching procedure or uncertainty over the matching procedure itself. The matching procedures that I consider are:

- nearest neighbor matching on the Mahalanobis distance
- nearest neighbor matching on the predictive mean (often used in the missing data imputation literature)
- nearest neighbor matching on the propensity score
- subclassification on the propensity score

While there are numerous matching procedures to consider, I focus on these four methods because they are relatively easy to estimate and understand, they allow for all observations to be matched, and they have been used extensively by researchers. For each matching procedure, I match N times, once for each observation in the data. I match with replacement in the sense that an observation can be a part of more than one of the N donor pools, but each observation may only be used once per pool. I also test each procedure using multiple donor pool sizes, varying the choice of donor pool size.

Although the authors of the various procedures have demonstrated the performance of their procedures in estimating average treatment effects, none of the procedures attain the ideal of exact matching. The procedures are simply a means to achieve covariate balance, where the distributions of the covariates are similar across treatment and control groups. In this case, since the comparison is between a single observation and a donor pool, the analogue to balance is simply whether the donor pool observations are exact matches to i. Deviations from exact matches creates bias in what is known as the matching discrepancy. Abadie and Imbens (2006) argue that the bias from the matching discrepancy may be negligible for ATEs when matching on a scalar or when the number of observations is large. It is unclear how the bias from the matching discrepancy affects the estimates of θ_i^{mis} and τ_i. Apart from the matching discrepancy, there may also be bias because of the estimation uncertainty around θ_i^{mis}, for which a Bayesian model accounts, as described below.

4.2 The Imputation Step

Since estimating τ_i is essentially a missing data problem where Y_i^{mis} is missing, the methods used are very similar to multiple imputation to deal with missing data (Rubin, 1987; Little and Rubin, 1987). Once I get an estimate of θ_i^{mis}, I need to fill in a value for Y_i^{mis}, denoted by $\tilde{Y}_i^{mis}$ to calculate $\tilde{\tau}_i$.[8] The imputation step is necessary to account for some fundamental uncertainty associated with $X_i^{(u)}$ that the matching does not account for. Y_i^{mis} should be imputed with values consistent with the observed Y values, so $\tilde{Y}_i^{mis}$ should be binary for binary Y and continuous for continuous Y. Recall that Y_i^{mis} was assumed to be drawn from some distribution $f(\cdot)$ conditional on observed covariates:

$$Y_i^{mis} \quad \sim \quad f(\cdot|\theta_i^{mis}, X_i, W_i)$$

For the imputation, I use a parametric approach that follows Rubin (2008) by drawing a value of $\tilde{Y}_i^{mis}$ from its posterior predictive distribution and repeating the process multiple times for each i. I end up with many imputed $\tilde{Y}_i^{mis}$ for each i, which forms a posterior predictive distribution that characterizes both estimation and fundamental uncertainty. I then use that posterior predictive distribution to calculate a posterior distribution for τ_i. The performance of parametric imputation likely depends on how accurately θ_i^{mis} is estimated as well as the size of the donor pool.

4.3 A Bayesian Model for Estimating τ_i

The general method I introduce is very simple with the following steps:

1. Choose a matching procedure.
2. For each i, use the matching procedure to create a donor pool.
3. Impute the missing potential outcome Y_i^{mis} using the donor pool and an assumed parametric distribution.
4. Calculate τ_i from the observed and imputed missing potential outcomes.
5. Repeat 2-4 for all i.
6. Repeat 1-5 many times for uncertainty.

[8] The $\sim$ above a parameter refers to a simulated draw of that parameter from its posterior distribution.

I incorporate these steps into a Bayesian model for a coherent and statistically principled framework. The Bayesian model also allows for inclusion of priors when qualitative knowledge exists on any specific observations, although in general, I use uniform priors so that the results approximate those that may be derived from a non-Bayesian framework. The Bayesian model accounts for both estimation and fundamental uncertainty using Markov Chain Monte Carlo (MCMC) methods to simulate from the posterior distribution of the parameters.

Let θ denote the vector of parameters to be estimated. At the most general level, θ includes the vector of θ_i^{mis}, parameters from the matching procedure which are denoted by $\theta_{\mathcal{M}}$, and possibly the choice of matching procedure, denoted by $\mathcal{M}$.[9] Although $\mathcal{M}$ is treated as a parameter, the data tells us nothing about $\mathcal{M}$ so the marginal posterior is equal to the prior for $\mathcal{M}$. $\mathcal{M}$ is simply included here as an option to reflect the researcher's uncertainty over the best or "correct" matching specification.

The typical Bayesian posterior is expressed as

$$p(\theta|Y, X, W) \propto p(Y|\theta, X, W)p(\theta)$$

Since W is independent of the potential outcomes through the ignorability assumption and X is independent of the potential outcomes conditional on θ_i^{mis}, I suppress W and X from the conditioning set for notational simplicity.

The idea behind this model is simple. Because the observed Y_i is fixed when i is the individual of interest, the only randomness comes from Y_i^{mis}. Nature randomly generates observations that come from the same distribution as Y_i^{mis}. The goal of the matching is to determine which of the observed observations comes from this distribution parameterized by θ_i^{mis}. The posterior is roughly translated into

$$p(\theta|Y, X, W) \propto \prod_{j=1}^{N} \{p(Y_j|\theta, X, W) \text{ given } j \text{ is a match for } i\} \times \text{priors}$$

The model is composed of two parts. The first part is a matching part to find the posterior for the parameters $\theta_{\mathcal{M}}$. The second part finds the posterior for θ^{mis}. I consider the matching part to be largely independent of the second part conditional on finding the observations matched. That is, once one knows which observations are matches, θ^{mis} is independent of $\theta_{\mathcal{M}}$. Depending on the matching procedure, the

[9] For example, $\mathcal{M}$ can be nearest neighbor 3-to-1 propensity score matching, in which case $\theta_{\mathcal{M}}$ are the coefficients in the propensity score equation. The researcher can vary $\mathcal{M}$ by choosing a different number of donor observations, changing how the distance metric is defined, or changing the set of matching variables.

matching parameters may or may not appear in the likelihood.[10] For simplicity and generality, I restrict my discussion of the likelihood term to simply focus on the likelihood for θ^{mis} assuming that the matching parameters are given.

4.3.1 Likelihood

The likelihood requires specifying the distribution that generated the data. Recall that our matching procedure is intended to generate a set of donor observations with "the same" values of X such that the donor observations are drawn from the same distribution as Y^{mis}. Now suppose one observes N binary variables $D^{(i)}$ (one variable for each i), which are indicators for whether j is a good match for i. Denote the set of $D^{(i)}$ variables as D. Then the likelihood[11] becomes

$$\begin{aligned}\mathcal{L}_{comp}(\theta^{mis}|Y,D) &= p(Y,D|\theta) \\ &= p(Y|D,\theta^{mis})p(D|\theta)\end{aligned}$$

This likelihood is known as the *complete data likelihood* as it refers to the likelihood if one were to observe the complete set of data including D. The distribution in the second term of the complete data likelihood is determined by matching:

$$p(D|\theta) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(D_j^{(i)}|\theta_{\mathcal{M}},\mathcal{M})$$

The first term in the complete data likelihood specifies the sampling distribution for the donor observations for each θ_i^{mis}:

$$\begin{aligned}p(Y|D,\theta^{mis}) &= \prod_{i=1}^{N}\prod_{j=1}^{N}\left[p(Y_j|\theta_i^{mis})\right]^{D_j^{(i)}} \\ &= \prod_{i=1}^{N}\prod_{j=1}^{N}\left[f(\cdot|\theta_i^{mis})\right]^{D_j^{(i)}}\end{aligned}$$

Since Y_i is assumed fixed and not modeled when estimating τ_i, this piece of the likelihood implies there is randomness only when an observation is used as a donor. The complete data likelihood is then rewritten

[10] For example, in Mahalanobis or propensity score matching, the outcome is not used so the matching parameters do not appear in the likelihood for Y. For predictive mean matching, the outcome is used. One can choose to model $\theta_{\mathcal{M}}$ separately or jointly with θ^{mis}. The process I describe models them separately by doing the matching independently first.

[11] Again I assume that the matching parameters are estimated separately and given.

as

$$\begin{aligned}\mathcal{L}_{comp}(\theta^{mis}|Y,D) &= \prod_{i=1}^{N}\prod_{j=1}^{N}\Big[p(Y_j|\theta_i^{mis})p(D_j^{(i)}|\theta_{\mathcal{M}},\mathcal{M}) + \\ &\quad p(Y_j|\theta_j^{other})\left(1-p(D_j^{(i)}|\theta_{\mathcal{M}},\mathcal{M})\right)\Big] \\ &= \prod_{i=1}^{N}\prod_{j=1}^{N}\Big[p(Y_j|\theta_i^{mis})p(D_j^{(i)}|\theta_{\mathcal{M}},\mathcal{M})\Big]^{D_j^{(i)}}\end{aligned}$$

In the first equation, θ_j^{other} simply refers to the fact that if j is not a match for i, then it is drawn from some other distribution that is not of interest. Therefore, the second term of the first equation drops out since non-matches do not contribute information to θ^{mis}. In all the likelihoods, the product over all i's indicates the full set of ICEs for every observation in the data.

The observed data likelihood simply integrates over our missing D:

$$\mathcal{L}_{obs}(\theta^{mis},Y) = \int p(Y,D|\theta^{mis})\,dD$$

The integral is generally mathematically intractable but one can simulate from the posterior via MCMC methods. The Bayesian model presented here uses the data augmentation algorithm of Tanner and Wong (1987). The original posterior is augmented with D to make computation more tractable.

4.3.2 Priors

All Bayesian models require specifying a prior distribution over all the parameters in the model. In this case, a prior is needed for θ_i^{mis}, $\theta_{\mathcal{M}}$, and $\mathcal{M}$. I assume that the parameters are independent a priori.

$$\begin{aligned}p(\theta) &= p(\theta^{(1)})p(\theta^{(2)})\ldots \\ &= p(\theta_{\mathcal{M}})p(\mathcal{M})p(\theta_i^{mis})\ldots p(\theta_N^{mis})\end{aligned}$$

For $\theta_{\mathcal{M}}$ and θ_i^{mis}, I generally use uninformative priors although one could incorporate qualitative knowledge into the priors. The choice of a prior for $\mathcal{M}$ boils down to which matching procedures one wants to consider. Since the data gives no information about the "best" matching procedure, the prior completely dominates the posterior for $\mathcal{M}$. If the researcher only wants to use one matching procedure as is typical in the causal inference literature, then the prior over $\mathcal{M}$ is essentially a spike prior. More research needs to be done on the influence of priors in my model on estimating individual causal effects.

4.3.3 Simulating from the Posterior via MCMC

I can simulate from the posterior of τ_i by using a Gibbs sampler, embedding the matching step within the sampler, and then drawing from the posterior predictive distribution (PPD) and calculating τ_i. For the Gibbs sampler, I draw from the full conditional distributions of the parameters conditional on the other parameters. The steps to simulate from the posterior of τ_i are:

MCMC Algorithm for the Posterior of τ_i

Repeat the following n_{sim} times:[a]

Gibbs Sampler:

1. Draw a matching procedure $\tilde{\mathcal{M}}$ from $p(\mathcal{M})$.
2. Draw $\tilde{\theta}_{\mathcal{M}}$ from $p(\theta_{\mathcal{M}}|Y, X, W, D, \theta^{mis}, \mathcal{M})$.

for (i in 1:N){

3. Determine $\tilde{D}^{(i)}$ from matching procedure. **(matching step)**
4. Draw $\tilde{\theta}_i^{mis}$ to estimate θ_i^{mis}.

}

Draw from PPD and Calculate τ_i:

for (i in 1:N){

5. Draw $\tilde{Y}_i^{mis}$ from $f(\cdot|\tilde{\theta}_i^{mis})$. **(imputation step)**
6. Calculate $\tilde{\tau}_i = W_i(Y_i - \tilde{Y}_i^{mis}) + (1 - W_i)(\tilde{Y}_i^{mis} - Y_i)$.

}

[a] Each draw of a parameter should be conditional on the current or previous draws of the other parameters. I have suppressed the iteration notation for aesthetic purposes.

1. $\mathcal{M}$ refers to any specification within the matching procedure. This can include any specification such as donor pool size, distance metric, or even the complete matching procedure itself. This leads to an important flexibility that my model allows, namely that I can simulate over the uncertainty of which matching procedure or which specifications within the matching procedure to choose. The data and other parameters do not generally give any information about model specification, so the full conditional is

$$p(\mathcal{M}|Y, X, W, \theta_{\mathcal{M}}, D, \theta^{mis}) = p(\mathcal{M})$$

which means that uncertainty over $\mathcal{M}$ is driven completely by the prior.[12] This flexibility is still

[12] The assumption that there is no information inherent in the data to distinguish between matching procedures is a simplifying assumption. One can imagine that the data provides information on which matching procedures are "better" by evaluating empirical balance in the covariates under each procedure and sampling the procedures probabilistically depending on the balance measure. More research into the feasibility of such approaches should be done.

useful in the case where the researcher is equally unsure about the various matching procedures and/or the number of observations in the donor pool, in which case he would put a uniform prior over the various permutations and incorporate that uncertainty within the simulation. In essence, allowing for uncertainty over $\mathcal{M}$ is similar to Bayesian model averaging approaches prevalent in the literature (Raftery, 1995; Montgomery and Nyhan, 2010). One important caveat is that $\mathcal{M}$ should produce a set of matches for the same individuals all the time or else the quantities of interest are unclear. The researcher may also simply choose to use one matching procedure, in which case $p(\mathcal{M})$ is a spike prior.

2. $\theta_{\mathcal{M}}$ represents possible parameters in the matching procedure. One example would be the coefficients in a model to estimate a propensity score or prognostic score. Not all matching procedures have parameters to be estimated, so step 2 may be skipped. The full conditional is

$$p(\theta_{\mathcal{M}}|Y, X, W, \mathcal{M}, D, \theta^{mis}) = p(\theta_{\mathcal{M}}|Y, X, W, \mathcal{M})$$

because $\theta_{\mathcal{M}}$ is estimated from the observed data and only depends on the data and the matching procedure used.

3. $D^{(i)}$ is calculated directly from the first two steps. $\mathcal{M}$ and $\theta_{\mathcal{M}}$ determine the rules by which an observation is considered a match so once $\mathcal{M}$ and $\theta_{\mathcal{M}}$ are known, D_i is completely determined. The other parameters do not affect $D^{(i)}$, so the full conditional can be thought of as

$$p(D^{(i)}|Y, X, W, \theta_{\mathcal{M}}, \mathcal{M}, \theta^{mis}) = p(D^{(i)}|Y, X, W, \theta_{\mathcal{M}}, \mathcal{M})$$

where the full conditional is a spike. Any uncertainty or randomness over $D^{(i)}$ is simply a function of uncertainty over $\mathcal{M}$ and/or $\theta_{\mathcal{M}}$. I also consider each $D^{(i)}$ to be independent so that an observation can be a donor for multiple donor pools.

4. θ_i^{mis} is finally estimated from the matched sample. Conditional on D_i, estimating θ_i^{mis} requires simply estimating the mean from a sample consisting of the donor pool. In most cases, if conjugate priors are chosen, then the full conditionals are also conjugates where

$$p(\theta_i^{mis}|Y, X, W, D, \theta_{\mathcal{M}}, \mathcal{M}) \quad = \quad p(\theta_i^{mis}|Y, D^{(i)})$$

What was previously an intractable posterior for θ_i^{mis} becomes incredibly easy to simulate from with the augmentation of D. Once the donor pool is known, it is simply a matter of modeling the donor pool. The draws of $\tilde{\theta}_i^{mis}$ form the posterior distribution of θ_i^{mis} and capture the estimation

uncertainty.

5. After simulating n_{sim} values of $\tilde{\theta}_i^{mis}$ from the posterior, I impute Y_i^{mis} by drawing one $\tilde{Y}_i^{mis}$ for each $\tilde{\theta}_i^{mis}$ from the posterior predictive distribution

$$p(Y_i^{mis}|Y) = \int p(Y_i^{mis}|\theta)p(\theta|Y)d\theta$$

Simply put, the model uses each draw of $\tilde{\theta}_i^{mis}$ and predicts a value of Y_i^{mis} by drawing from $f(\cdot|\tilde{\theta}_i^{mis})$. While the estimation uncertainty is captured by the n_{sim} draws of $\tilde{\theta}_i^{mis}$, the fundamental uncertainty is captured by the sampling in this step.

6. Drawing from the posterior of τ_i is straightforward given that there is a deterministic relationship between τ_i, Y_i, and Y_i^{mis}. Let the posterior distribution for τ_i be

$$p(\tau_i|Y) = \int p(\tau_i|Y_i^{mis}, Y)p(Y_i^{mis}|Y)dY_i^{mis}$$

where $p(\tau_i|Y_i^{mis}, Y)$ is a spike distribution. Since I have simulations from $p(Y_i^{mis}|Y)$, the posterior of τ_i can be simulated simply by taking each draw of $\tilde{Y}_i^{mis}$ and calculating

$$\tilde{\tau}_i = W_i(Y_i - \tilde{Y}_i^{mis}) + (1 - W_i)(\tilde{Y}_i^{mis} - Y_i)$$

Note that in the algorithm, steps 3-6 are conducted separately for each i. Although in practice, the steps may be done altogether for all i, I choose to characterize the i's separately for both pedagogical and substantive purposes. One should consider each τ_i as a separate estimand estimated separately to avoid criticisms of multiple testing and cherry-picking specific ICEs. Theoretically, one should think of this framework as conducting N separate studies to estimate N different causal effects. For each study, imagine a dataset consisting only of observation i and all observations j where $j \neq i$ and $W_j = W_i$. In this framework, each observation may be used as a donor observation for multiple pools. When estimating ATEs, researchers who match with replacement must reweight the donor observations to reflect the correct number of observations in the data. In the case of estimating ICEs, no reweighting is necessary from a conceptual standpoint since the N ICEs are estimated in "separate" studies. However, if certain observations are used as donors many times, the multiple testing problem may be exacerbated, especially if the repeat donors are outliers. Overall, it is still unclear how including observations in multiple donor pools affects the estimates of the variances of the ICEs.

4.4 Comparison to Existing Approaches

I see a few contributions of the framework and model I have proposed. In addition to calling attention to focusing on ICEs in general, my model combines the ideas of matching and Bayesian analysis to estimate different causal quantities of interest. My model is flexible in the choice of matching and also allows for exploration and discovery of different treatment effects and treatment effect heterogeneity.

The approach I use to estimate ICEs bears many similarities to existing frameworks. I now discuss the similarities between my approach and the approach laid out by Rubin first in Rubin (1978) and then discussed in Rubin (2008) and most recently extended in Pattanayak, Rubin and Zell (2012), hereafter known as PRZ.[13] While none of the papers explicitly discuss individual causal effects as a quantity of interest, they all allow for the imputation of missing potential outcomes using Bayesian methods, which is also a characteristic of my approach. I argue that although there are subtle differences between my approach and the Rubin approach, my framework can be described as a generalization of the Rubin framework.

The first difference between the two approaches is in the data generating process and defining what is random. The Rubin approach assumes that Y, W, and X are all realizations from random variables whereas I assume that W and X are fixed and Y is only random because of unmeasured prognostic variables. I see the distinction between the two approaches on this point to be negligible. The idea of unmeasured prognostic variables leading to random outcomes is not incompatible with the Rubin approach. Furthermore, both approaches place great importance on the assumptions of ignorability of treatment assignment and SUTVA. My approach also allows for conditioning on non-confounding prognostic variables to improve the imputations of the missing potential outcomes. Since the estimand in the Rubin approach is an average treatment effect, including non-confounding prognostic variables is less important although in many cases, it can lead to more efficient estimates.

A second difference between the two approaches is that the Rubin approach models the observed outcomes whereas I keep the observed outcomes fixed. On the surface, this may seem like a big difference. But in reality, the difference is mostly in the framing of the problem rather than any substantive differences. The Rubin approach estimates θ_T and θ_C, which are the means of the distributions of treated and control units, from the observed treated and observed control units respectively. PRZ go one step further by stratifying observations either by their propensity scores or by existing substantive strata and

[13] The PRZ approach has a very specific model and specific quantities of interest that are applicable to their data and question. I describe the PRZ approach in very general terms and discuss how the general PRZ setup compares to my framework.

estimating a separate pair of θ for each strata. The Rubin approach then draws the missing potential outcomes from distributions centered at θ_T and θ_C. This is exactly the same approach that I use. For a missing $Y_i(0)$ outcome, θ_i^{mis} is estimated from a donor pool of control observations deemed to be good matches. Similarly, for a missing $Y_i(1)$ outcome, θ_i^{mis} is estimated from a donor pool of treated observations deemed to be good matches.

The difference is that each observation has a separate θ_i^{mis} to impute its missing potential outcome. In the earlier Rubin approaches, there are only two θ's, a θ_C to impute for treated units and a θ_T to impute for control units. PRZ allows for more flexibility by having strata-specific θ's. My approach basically generalizes PRZ by allowing each i to have its own individual strata.

To see this more clearly, suppose that there is a strata consisting of two treatment units, T_1 and T_2, and two control units C_1 and C_2. All four units are deemed to be good matches for each other, so assume ignorability of treatment assignment. In the PRZ approach, one would impute the missing $Y(0)$ for T_1 and T_2 with $\tilde{\theta}_C$ estimated from C_1 and C_2. Similarly, one would impute the missing $Y(1)$ for C_1 and C_2 with $\tilde{\theta}_T$ estimated from T_1 and T_2. Under my approach, the missing outcome for T_1 is imputed from $\tilde{\theta}_{T_1}$ estimated from C_1 and C_2, the missing outcome for T_2 is also imputed from the same $\tilde{\theta}_{T_2}$ estimated from C_1 and C_2 where $\tilde{\theta}_{T_1} = \tilde{\theta}_{T_2}$ and the missing outcomes for C_1 and C_2 are imputed from the $\tilde{\theta}_{C_1}$ and $\tilde{\theta}_{C_2}$ estimated from T_1 and T_2, where $\tilde{\theta}_{C_1} = \tilde{\theta}_{C_2}$. The two approaches are exactly the same assuming that my matching procedure produces the same strata. However, my approach is more generalizable in that the researcher can implement a matching procedure that does not restrict the donor pool to be within the same strata. T_1 can have a donor pool of C_1 and C_2 whereas T_2 can have a donor pool of C_1, C_2, and C_3.

This brings us to a third difference between my approach and the existing Rubin approach, namely that my framework allows for matching and uncertainty in the matching procedure and matching parameters. In PRZ, the strata are assumed to be exogenously defined or estimated beforehand with propensity score stratification. Once the strata are defined, they cannot be changed and the donor pool stays constant. My approach allows for multiple matching procedures and uncertainty within each matching procedure to characterize uncertainty about which observations constitute the correct donor pools. In approximate matching methods, this uncertainty certainly exists amongst researchers.

The Bayesian model I have proposed is unique in a couple ways. First, my model is explicit in that the quantity of interest is the individual causal effects. Most models estimate average treatment effects and consider the individual effects only indirectly if at all. Second, the data generating process I propose is

slightly unconventional. Third, it embeds a relatively non-parametric matching step in the imputation of Y^{mis}. Finally, it allows for uncertainty over parameters within the matching procedure or uncertainty over which matching procedure to choose itself. As with any Bayesian model, the model is sensitive to choice of priors and convergence is not guaranteed in finite time. However, the ability to use priors also has the advantage of incorporating substantive information or restricting the range of possible values to help overcome sample size issues.

5 Other Quantities of Interest

Once the posterior for the individual causal effects is obtained, any sample estimand can be calculated rather easily by aggregating subsets of individual causal effects. For example, the posterior of the sample average treatment effect can be obtained by averaging the set of draws of τ_i for all i at each iteration of the Markov chain. Similarly, my approach allows for discovery and exploration of treatment effect heterogeneity by averaging over subsets of τ_i, such as averaging the draws for the τ_i for treated individuals to get the posterior of the sample ATT, averaging over draws for subsets of individuals with certain covariate values to get the sample CATE, etc. The researcher can graphically visualize heterogeneity by plotting the ICEs against various covariates. One can also ask questions such as the probability that the sample CATE is greater for individuals with $X = a$ versus individuals with $X = b$ for any values a and b simply by differencing the posterior draws. Obtaining posterior draws for τ_i for every individual in the sample allows for almost limitless possibilities to examine treatment effect heterogeneity.

Although various sample estimands are easy to calculate with this framework, it is unclear how one would estimate population or super-population estimands under my framework. Recall that the model assumes a finite sample and a Bayesian framework. It imputes the missing potential outcome for each individual in the sample while allowing the observed outcome to be fixed and unmodeled. The framework does not extend easily to super-population estimands because both potential outcomes are missing for individuals not in the sample. One way to get at super-population estimands may be to use bootstrapping. For each bootstrapped sample, calculate the estimands using the estimated posterior for the bootstrapped individuals and repeat to obtain a posterior over the super-population estimand. However, this process assumes that our sample is completely representative of the super-population. More specifically, it assumes that every other individual not observed in the super-population is exactly the same as an individual in our observed dataset. Furthermore, the bootstrap process almost certainly underestimates the uncertainty around super-population estimates because of the fixed and unmodeled

potential outcome in the model. Generally speaking, the idea of estimating individual causal effects and estimating super-population estimands are contradictory in the sense that a super-population by definition contains nameless and exchangeable individuals whereas individual causal effects involve specific individuals in the dataset. For these reasons, I restrict the framework to estimating sample estimands of interest.

Another related issue is whether or not my framework allows for out-of-sample predictions or predictions for future observations. For reasons similar to those for estimating super-population effects, out-of-sample predictions are not straightforward. One can reasonably predict the treatment effect for an out-of-sample observation by finding and using the results for an in-sample observation with a similar covariate profile. For out-of-sample observations where no in-sample observations match reasonably well, the data does not give much information and a parametric model is needed. However, I argue that the same issues of model dependence for prediction occur in any other estimation framework. My model uses all available information in the data.

6 Applications and Extensions

The framework I have introduced is flexible enough to be applied to many situations and can be extended in various ways. Some applications and extensions to consider include:

- **Binary treatment with any type of outcome variable**: The simplest situation that I apply the model to is a dataset with a binary treatment variable, various covariates, and an outcome variable of any type. The outcome can be continuous or discrete, and treatment should be ignorable given the observed covariates.
- **Non-binary treatment**: The framework can be easily extended to non-binary treatment variables by retaining a linearity assumption. Instead of two potential outcomes, each individual has possibly an infinite number of potential outcomes. However, by assuming a linear relationship between the treatment and the outcome, one only needs to impute one missing potential outcome and extrapolate the rest by assumption. The linearity assumption also allows researchers to use individuals with significantly different treatment values to impute the same missing potential outcome.
- **Missing data in the covariates**: Since the model uses a Bayesian framework, one can easily incorporate imputation of missing data in the covariates via any of the existing multiple imputation techniques prevalent in the missing data literature.

- **Two-stage models**: The two related topics of treatment non-compliance and instrumental variables can be incorporated into the model via existing techniques. For example, one can model treatment non-compliance via principal stratification (Frangakis and Rubin, 2002) by applying ICEs into the first stage of a two-stage model and incorporating existing Bayesian models (Imbens and Rubin, 1997) into the sampler. The researcher can then use the principal stratifications from the first stage to calculate ICEs in the second stage. The framework can also be used to test the monotonocity assumption in instrumental variables models by estimating individual causal effects in the first stage.

- **Time-series cross-sectional/panel/multiple measurements data**: The framework can also handle data where individuals are measured repeatedly over time. Multiple measurements of outcomes and/or covariates and treatment give the researcher more information to match on and impute with. One would simply need to model the time component and decide on how to incorporate the extra information into the framework.

The next chapter tests various aspects of my framework via simulation to see how well various methods can recover individual causal effects. I then present various applications of my framework to real data and questions of interest to academics and policymakers in the general social science world.

References

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105(490):493–505.

Abadie, Alberto and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74(1):235–267.

An, Weihua. 2010. "Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference." *Sociological Methodology* 40(1):151–189.

Arceneaux, Kevin and David W. Nickerson. 2009. "Who Is Mobilized to Vote? A Re-Analysis of 11 Field Experiments." *American Journal of Political Science* 53(1):1–16.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens and Oscar A. Mitnik. 2008. "Nonparametric Tests for Treatment Effect Heterogeneity." *The Review of Economics and Statistics* 90(3):389–405.

Diamond, Alexis and Jasjeet S. Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *Review of Economics and Statistics* . Forthcoming.

Dorresteijn, Johannes A., Frank L. Visseren, Paul M. Ridker, Annemarie M. Wassink, Nina P. Paynter, Ewout W. Steyerberg, Yolanda van der Graaf and Nancy R. Cook. 2011. "Estimating Treatment Effects for Individual Patients Based on the Results of Randomised Clinical Trials." *British Medical Journal* 343:d5888.

Feller, Avi and Chris C. Holmes. 2009. "Beyond Toplines: Heterogeneous Treatment Effects in Randomized Experiments." Unpublished.

Frangakis, Constantine E. and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58(1):21–29.

Gadbury, Gary L., Hari K. Iyer and Jeffrey M. Albert. 2004. "Individual Treatment Effects in Randomized Trials with Binary Outcomes." *Journal of Statistical Planning and Inference* 121(1):163–174.

Gaines, Brian J. and James H. Kuklinski. 2011. "Experimental Estimation of Heterogeneous Treatment Effects Related to Self-Selection." *American Journal of Political Science* 55(3):724–736.

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Gutman, Roee and Donald B. Rubin. 2012. "Robust Estimation of Causal Effects of Binary Treatments in Unconfounded Studies with Dichotomous Outcomes." *Statistics in Medicine* .

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20(1):25–46.

Hansen, Ben B. 2008. "The Prognostic Analogue of the Propensity Score." *Biometrika* 95(2):481–488.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition ed. New York: Springer.

Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199–236.

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

Iacus, Stefano M., Gary King and Giuseppe Porro. 2012. "Causal Inference without Balance Checking: Coarsened Exact Matching." *Political Analysis* 20(1):1–24.

Imai, Kosuke and Aaron Strauss. 2011. "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-out-the-vote Campaign." *Political Analysis* 19(1):1–19.

Imai, Kosuke, Gary King and Elizabeth A. Stuart. 2008. "Misunderstanding Between Experimentalists and Observationalists About Causal Inference." *Journal of the Royal Statistical Society Series A* 171(2):481–502.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *Annals of Applied Statistics* . Forthcoming.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87(3):706–710.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *The Review of Economics and Statistics* 86(1):4–29.

Imbens, Guido W. and Donald B. Rubin. 1997. "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *The Annals of Statistics* 25(1):305–327.

Jin, Hui and Donald B. Rubin. 2008. "Principal Stratification for Causal Inference with Extended Partial Compliance." *Journal of the American Statistical Association* 103(481):101–111.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):341–355.

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton, NJ: Princeton University Press.

Kravitz, Richard L., Naihua Duan and Joel Braslow. 2004. "Evidence-Based Medicine, Heterogeneity of Treatment Effects and the Trouble with Averages." *The Millbank Quarterly* 82(4):661–687.

Lagakos, Stephen W. 2006. "The Challenge of Subgroup Analyses - Reporting without Distorting." *The New England Journal of Medicine* 354(16):1667–1669.

Little, Roderick J. A. and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data.* New York: Wiley.

McCullagh, Peter and John A. Nelder. 1989. *Generalized Linear Models.* Second edition ed. New York: Chapman and Hall.

Ming, Kewei and Paul R. Rosenbaum. 2000. "Substantial Gains in Bias Reduction from Matching with a Variable Number of Controls." *Biometrics* 56(1):118–124.

Montgomery, Jacob M. and Brendan Nyhan. 2010. "Bayesian Model Averaging: Theoretical Developments and Practical Applications." *Political Analysis* 18(2):245–270.

Pattanayak, Cassandra W., Donald B. Rubin and Elizabeth R. Zell. 2012. "A Potential Outcomes, and Typically More Powerful, Alternative to "Cochran-Mantel-Haenszel"." Working Paper.

Pocock, Stuart J., Susan E. Assmann, Laura E. Enos and Linda E. Kasten. 2002. "Subgroup Analysis, Covariate Adjustment and Baseline Comparisons in Clinical Trial Reporting: Current Practice and Problems." *Statistics in Medicine* 21(19):2917–2930.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111–163.

Rosenbaum, Paul R. 1984. "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society Series A* 147(5):656–666.

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

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. "The Bias Due to Incomplete Matching." *Biometrics* 41(1):103–116.

Rothwell, Peter M. 2005. "Subgroup Analysis in Randomised Controlled Trials: Importance, Indications and Interpretation." *The Lancet* 365(9454):176–186.

Rubin, Donald B. 1973*a*. "Matching to Remove Bias in Observational Studies." *Biometrics* 29(1):159–183.

Rubin, Donald B. 1973*b*. "The Use of Matching Sampling and Regression Adjustment to Remove Bias in Observational Studies." *Biometrics* 29(1):185–203.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688–701.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6(1):34–58.

Rubin, Donald B. 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74(366):318–328.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100(469):322–331.

Rubin, Donald B. 2008. Statistical Inference for Causal Effects, With Emphasis on Applications in Epidemiology and Medical Statistics. In *Handbook of Statistics*, ed. C.R. Rao, J. Philip Miller and D.C. Rao. Vol. 27 Elsevier pp. 28–63.

Rubin, Donald B. and Neal Thomas. 2000. "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association* 95(450):573–585.

Rubin, Donald B. and Richard P. Waterman. 2006. "Estimating the Causal Effects of Marketing Interventions Using Propensity Score." *Statistical Science* 21(2):206–222.

Steyer, Rolf. 2005. "Analyzing Individual and Average Causal Effects via Structural Equation Models." *Methodology* 1(1):39–54.

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25(1):1–21.

Tanner, Martin A. and Wing Hung Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82(398):528–540.