TECHNICAL WORKING PAPER SERIES

USING STUDIES OF TREATMENT RESPONSE TO INFORM TREATMENT CHOICE IN HETEROGENEOUS POPULATIONS

Charles F. Manski

Technical Working Paper 263
http://www.nber.org/papers/T0263

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2000

This research was supported in part by National Science Foundation grant SES-0001436. A draft of part of the content of Section 4 was circulated in May 1999 as "Statistical Treatment Rules for Heterogeneous Populations." The author has benefitted from the comments of John Pepper and the opportunity to present this work in seminars at Johns Hopkins University, Eli Lilly & Co., New York University, Northwestern University, the University of Bristol, the University of Chicago, and the University of Virginia. The author thanks William G. Henderson, Domenic Reda, and David Williams of the Edward Hines, Jr., Hospital, U.S. Department of Veterans Affairs Cooperative Studies Program Coordinating Center, Hines IL, for providing and helping to interpret the clinical-trial data analyzed in Section 3. The Department of Veterans Affairs retains ownership of the data. The views expressed in this paper are those of the author and are not necessarily endorsed by the Department of Veterans Affairs or the National Bureau of Economic Research.

Using Studies of Treatment Response to Inform Treatment Choice
in Heterogeneous Poplulations
Charles F. Manski
NBER Technical Working Paper No. 263
September 2000
JEL No. C10, D81

ABSTRACT

An important practical objective of empirical studies of treatment response is to provide decision makers with information useful in choosing treatments. Often the decision maker is a planner who must choose treatments for the members of a heterogeneous population; for example, a physician may choose medical treatments for a population of patients. Studies of treatment response cannot provide all the information that planners would like to have as they choose treatments, but researchers can be of service by addressing several questions: How should studies be designed in order to be most informative? How should studies report their findings so as to be most useful in decision making? How should planners utilize the information that studies provide? This paper addresses aspects of these broad questions, focusing on pervasive problems of identification and statistical inference that arise when studying treatment response.

Professor Charles F. Manski
Department of Economics
Northwestern University
2003 Sheridan Road
Evanston, IL 60208
and NBER
cfmanski@northwestern.edu

1. Introduction

An important practical objective of empirical studies of treatment response is to provide decision makers with information useful in choosing treatments. Often the decision maker is a *planner* who must chooses treatments for a heterogeneous population. The planner might, for example, be a physician choosing medical treatments for a population of patients. Physicians use findings of medical research to evaluate the merits of alternative treatment rules.

It is unrealistic to think that studies of treatment response can provide all the information that planners such as physicians would like to have as they choose treatments. Some divergence between the conditions of studies and of actual decision problems seems inevitable. However, researchers can aim to improve treatment choice by addressing several questions: How should studies be designed in order to be most informative? How should studies report their findings so as to be most useful in decision making? How should planners utilize the information that studies provide? This paper addresses aspects of these broad questions, drawing in part on the author's past research and in part on new ideas developed here.

My starting point is the decision theoretic framework of Manski (2000). This assumes that the planner observes some covariates for each member of the population to be treated; for example, a physician may observe a patient's demographic attributes, medical history, and the results of diagnostic tests. The observed covariates determine the set of non-randomized treatment rules that are feasible for the planner to implement: the set of feasible rules is the set of all functions mapping the observed covariates into treatments. Each member of the population has a response function which maps treatments into a real-valued outcome of interest; perhaps a measure of health status in the case of medical treatment. I assume that the planner wants to choose a treatment rule that maximizes the population mean outcome; in economic terms, the planner wants to maximize a utilitarian social welfare function. Under these assumptions, an optimal treatment rule assigns to each member of the population a treatment that maximizes mean outcome conditional on the person's observed covariates. Hence studies of treatment response are useful to the degree that they enable the planner to learn how mean outcomes vary with treatments and covariates.

Section 2 formalizes these ideas, from which I conclude that heterogeneity in treatment response should be a central concern in study design. In particular, researchers should bear in mind the planner's problem when deciding what population to study and what covariate information to collect and report on study subjects. I observe that research articles commonly provide only a small part of the covariate information that planners may find useful. I reconsider the widely held view that studies of treatment response should be judged primarily by their internal validity and only secondarily by their external validity. In so doing, I contrast the treatment-choice perspective advanced here with research seeking to learn "causal effects."

Sections 3 and 4 examine the implications for treatment choice of pervasive inferential problems in studies of treatment response, with Section 3 focusing on identification and Section 4 on statistical inference. Section 3 extends my past research on identification. I first explain how identification problems in the empirical analysis of treatment response generate ambiguity about the identity of optimal treatment choices. I then ask how a planner choosing a treatment rule should approach the fundamental *selection problem* that occurs when one attempts to interpret data on treatment response in the absence of knowledge of the process of treatment selection. I examine the different forms taken by the selection problem in observational studies and in randomized experiments with partial compliance. I reiterate the point first made in Manski (1990) that observational studies are informative even when nothing is known about the process of treatment selection, but not sufficiently informative to resolve ambiguity in treatment choice. Considering experiments with partial compliance, I explain how the inferential problem depends on whether noncomplying subjects can "cross over" to other treatments under study. Section 3 concludes by expanding the discussion from the selection problem, which manifests itself as missing outcome data, to consideration of treatment choice using studies with missing covariate and outcome data. Here I summarize and apply findings in Horowitz and Manski (1998, 2000).

Studies of treatment response generally report the outcomes of samples of subjects, not of entire study populations. Hence planners seeking to learn from research face not only the deductive identification problems addressed in Section 3 but also inductive statistical problems of inference from samples to populations. Section 4 first sets out general principles for evaluation of *statistical treatment rules*. I then give these principles content by examining two familiar frequentist rules in classical experimentation settings, where mean treatment response is identified. The *conditional success* (*CS*) rule and the *unconditional success* (*US*) rule both embody the reasonable idea that persons should be assigned the treatment with the best empirical success rate, but they differ in their use of covariate information; the CS rule selects treatments with the best empirical success rates conditional on specified covariates, and the US rule selects a treatment with the best unconditional empirical success rate. Whereas the US Rule constrains the planner to choose the same treatment for all persons, the CS Rule permits the planner to treat persons with different covariates differentially. Whereas the US Rule has the planner compare success rates using the entire available sample, the CS Rule requires that the planner compare success rates in sub-samples. There is an evident tension between use of covariate information and available sample size. I characterize this tension and assess the implications for treatment choice. Section 4 concludes by drawing implications for study design and for the interpretation of research findings. I contrast the treatment-choice perspective on statistical inference with the hypothesis-testing perspective that has long prevailed in research on treatment response.

In each of Sections 3 and 4, I use a clinical trial of treatments for hypertension (Materson *et al.*, 1993; Materson, Reda, and Cushman, 1995) to illustrate findings. Heterogeneity in treatment response has been an important theme of medical research on hypertension, which holds that "Antihypertensive treatment must be tailored to the individual patient" (Materson *et al.*, 1993, p. 919.) In Section 3, I elaborate on the empirical analysis of Horowitz and Manski (2000), which used data from the hypertension trial to show how missing covariate and outcome data may affect inference in randomized experiments. In Section 4, I use the trial to demonstrate how a physician might choose between the CS and US rules. Medical illustrations such as the hypertension trial are particularly salient because the idea that research should aim to inform treatment choice is well-developed in medicine. In the United States, the Food and Drug Administration (FDA) does not approve the marketing of new drugs until their efficacy has been established in clinical trials. Health insurance providers commonly refuse to cover new medical procedures until sufficient evidence on their efficacy accumulates.

Of course medicine is hardly the sole application for the analysis of this paper. A judge deciding sentences for a population of convicted offenders is a planner making treatment choices (Manski and Nagin, 1998). So is a case manager placing welfare recipients in employment training programs (Doolittle and Riccio, 1992; Pepper, 2000). Criminological research on sentencing and recidivism may provide information useful to judges and research evaluating social programs may inform welfare case managers. The analysis that follows should be relevant in these settings and many more.

2. Treatment Choice in Heterogeneous Populations

Section 2.1 formalizes the planner's problem as in Manski (2000). With this background, Sections 2.2 and 2.3 draw implications for studies of treatment response.

2.1. The Planner's Problem

2.1.1. The Choice Set and the Objective Function

I suppose that there is a finite set T of mutually exclusive and exhaustive treatments. A planner must choose a treatment rule assigning a treatment in T to each member of a population J. Each person $j \in J$ has a *response function* $y_j(\cdot)$: T → Y mapping treatments into real-valued outcomes $y_j(t) \in Y$. A *treatment rule* is a function $\tau(\cdot)$: J → T specifying which treatment each person is assigned. Thus person j's outcome under rule $\tau(\cdot)$ is $y_j[\tau(j)]$. This notation maintains the assumption of individualistic treatment made commonly in analyses of treatment response. That is, a person's outcome may depend on the treatment he is assigned, but not on the treatments assigned to others.

The planner is concerned with the distribution of outcomes across the population, not with the outcomes of particular persons. Hence the population is taken to be a probability space, say (J, Ω, P), where Ω is the σ-algebra on which probabilities are defined and P is the probability measure. Now the population mean outcome, or *social welfare*, under treatment rule $\tau(\cdot)$ is well-defined as

$$(1)\quad E\{y_j[\tau(j)]\} \equiv \int y_j[\tau(j)]dP(j).$$

I assume that the planner wants to choose a treatment rule that maximizes $E\{y_j[\tau(j)]\}$.

This criterion function has normative, analytical, and practical appeal. Maximization of a population mean outcome, or perhaps some weighted average outcome, is the standard normative criterion of the public economics literature on social planning; the outcome of interest measures the social benefits minus costs of a treatment. The linearity of the expectation operator yields substantial analytical simplifications, particularly through use of the law of iterated expectations. The practical appeal is that a planner choosing treatments to maximize the mean population outcome will want to learn average treatment effects, the dominant form of treatment effect reported in the empirical literature on treatment response. Other criterion functions generate interest in other forms of treatment effect. For example, Heckman, Clements, and Smith (1997) consider criterion functions that incorporate equity considerations.

The planner observes certain covariates $x_j \in X$ for each member of the population. The planner cannot distinguish among persons with the same observed covariates and so cannot implement treatment rules that systematically differentiate among these persons. Hence the feasible non-randomized rules are functions mapping the observed covariates into treatments. I do not explicitly consider randomized treatment rules, but there is a simple implicit way to permit such rules. Let x include a component whose value is randomly drawn by the planner from some distribution. Then the planner can make the chosen treatment vary with this covariate component.

To formalize the planner's problem, let Z denote the space of all functions mapping X into T. Let $z(\cdot) \in Z$. Then the feasible treatment rules have the form

$$(2) \quad \tau(j) = z(x_j), \qquad j \in J.$$

Let $P[y(\cdot), x]$ be the probability measure on $Y^T \times X$ induced by P(j). Let $E\{y[z(x)]\} \equiv \int y[z(x)]dP[y(\cdot), x]$ denote the expected value of y[z(x)] with respect to this induced measure. Then the planner wants to solve the problem

$$(3) \quad \max_{z(\cdot) \in Z} \; E\{y[z(x)]\}.$$

In practice, institutional constraints may restrict the feasible treatment rules to some proper subset of the space Z. In particular, the planner may be precluded from using certain covariates (say race or gender) to assign treatments. The analysis in this paper continues to hold if x is defined to be the covariates that the planner is permitted to consider, rather than the full vector of covariates that the planner observes.

2.1.2. Optimal Treatment Rules and the Value of Covariate Information

It is easy to show that the solution to the planner's problem is to assign to each member of the population a treatment that maximizes mean outcome conditional on the person's observed covariates. Let

1[·] be the indicator function taking the value one if the logical condition in the brackets holds and the value zero otherwise. For each z(·) ∈ Z, use the law of iterated expectations to write

$$(4)\quad E\{y[z(x)]\} \;=\; E\{E\{y[z(x)]\,|\,x\}\} \;=\; E\{\sum_{t\in T} E[y(t)\,|\,x]\cdot 1[z(x)=t]\} \;=\; \int \sum_{t\in T} E[y(t)\,|\,x]\cdot 1[z(x)=t]\; dP(x).$$

For each x ∈ X, the integrand $\Sigma_{t\in T}$ E[y(t)|x]·1[z(x) = t] is maximized by choosing z(x) to maximize E[y(t)|x] on t ∈ T. Hence a treatment rule $z^*(\cdot)$ is optimal if , for each x ∈ X, $z^*(x)$ solves the problem

$$(5)\quad \max_{t\in T} E[y(t)\,|\,x].$$

The optimized population mean outcome is $E\{\max_{t\in T} E[y(t)\,|\,x]\}$.

The set of feasible treatment rules grows as more covariates are observed. Hence the optimal mean outcome achievable by the planner cannot fall, and may rise, as more covariates are observed. The *value of covariate information* is appropriately measured by the difference between the optimal mean outcome achievable with and without use of this information. This is

$$(6)\quad V(X) \;\equiv\; E\{\max_{t\in T} E[y(t)\,|\,x]\} \;-\; \max_{t\in T} E[y(t)].$$

Inspection of (6) shows that covariate information has no value if there exists a common optimal treatment; that is, a t* ∈ T such that z*(x) = t*, almost everywhere on X. Covariate information does have value if optimal treatments vary with x.

More generally, we may compare the value of observing distinct covariate vectors, say x and w. A planner who knows the conditional mean treatment responses E[y(·)|x] and E[y(·)|w] should prefer observation of x to w if and only if $E\{\max_{t \in T} E[y(t)|x]\} \geq E\{\max_{t \in T} E[y(t)|w]\}$. In words, the planner should prefer x to w if x better separates persons who differ in their optimal treatments.

Note that this criterion for comparison of x and w differs from the prediction criterion familiar in statistical decision theory. The prediction criterion supposes that, for each $t \in T$, one wants to predict y(t) as well as possible in the sense of minimizing expected square loss. The best predictors conditional on x and w are $E[y(t)|x]$ and $E[y(t)|w]$ respectively. A statistician who knows $E[y(t)|x]$ and $E[y(t)|w]$ and wants to predict y(t) as well as possible should prefer x to w if and only if $E\{y(t) - E[y(t)|x]\}^2 \leq E\{y(t) - E[y(t)|w]\}^2$.

2.2. Collection and Reporting of Covariate Information in Studies of Treatment Response

The above formalization of the planner's problem makes clear that heterogeneity of treatment response should be a central concern in study design. Perhaps the most immediate implication is that researchers should bear in mind the planner's problem when deciding what covariate information to collect and report on study subjects. Our discussion of the value of covariate information shows that such information is useful if the planner can observe the covariates in question and if optimal treatments vary across persons with different values of these covariates. The more that treatment response varies with covariates, the more valuable covariate information is to the planner.

Examination of medical research shows that there is often a wide disparity between the covariates that planners can observe and the covariate information reported in studies of treatment response. Physicians commonly observe extensive covariate data – medical histories, diagnostic test findings, and demographic attributes – for the patients that they treat. Yet the medical journal articles that report on clinical trials usually provide scant covariate information for the study subjects, describing outcomes only within broad risk-factor groups.

A recent article on a clinical trial comparing alternative psychosocial treatments for cocaine dependence provides an apt illustration. Crits-Christoph, P. Siqueland, L., Blaine, J., Frank, A., et. al. (1999) report on a National Institute on Drug Abuse study randomly placing 487 cocaine-dependent patients in one of four treatment groups, each designated treatment combining group drug counseling (GDC) with another form of therapy. In some respects, the article is attentive to the possibility of heterogeneity in treatment response. The authors call attention to the fact that previous findings on the relative merits of psychotherapy and drug counseling for treatment of opiate-dependent patients do not hold up in the context of cocaine dependence. They provide much descriptive information on the characteristics of the subjects; including measures of race, sex, age, education, employment status, type and severity of drug use, psychiatric state, and personality. They test hypotheses that treatment effects do not vary with patient psychiatric state or personality. However, the article does not report outcomes conditional on any of the patient covariates observed by the researchers. Indeed, its formal Conclusion section makes no reference to the possibility that treatment response might vary with covariates, stating simply (page 493): "Compared with professional psychotherapy, a manual-guided combination of intensive individual drug counseling and GDC has promise for the treatment of cocaine dependence."

There seem to be several reasons why researchers collect and report little covariate information. (I say "seem to" because these reasons are rarely stated explicitly.) Sometimes researchers seem to assume that there exists a common optimal treatment across the population of interest; then covariate information has no value (see Section 2.1.2). Sometimes concern for the confidentiality of subjects' identities inhibits researchers from collecting and reporting covariates that may be related to treatment outcomes. Sometimes editorial restrictions on the lengths of journal articles prevent researchers from reporting all potentially useful findings. Sometimes sampling variability inhibits researchers from reporting treatment response conditional on covariates; findings may be reported only if they meet conventional criteria for statistical precision (see Section 4.3 for further discussion). I think it would be better if researchers would collect extensive covariate information whenever there is reason to think that treatment response is heterogeneous. Subject to considerations of subject confidentiality, researchers should report mean treatment response conditional on these covariates, or at least make this information available on the web or through other means.

2.3. The Study Population and the Population to be Treated

A longstanding issue in study design concerns the importance of correspondence between the study population and the population to be treated. This matter was downplayed in the influential work of Donald Campbell, who long argued that studies of treatment effects should be judged primarily by their internal validity and only secondarily by their external validity (e.g., Campbell and Stanley, 1963; Campbell, 1984). Campbell's view has recently been endorsed by Rosenbaum (1999), who recommends that observational studies of human subjects aim to approximate the conditions of laboratory experiments (page 263):

> "In a well-conducted laboratory experiment one of the rarest of things happens: The effects caused by treatments are seen with clarity. Observational studies of the effects of treatments on human populations lack this level of control but the goal is the same. Broad theories are examined in narrow, focused, controlled circumstances."

Rosenbaum, like Campbell, downplays the importance of having the study population be similar to the population of interest, writing (page 259): "Studies of samples that are representative of populations may be quite useful in describing those populations, but may be ill-suited to inferences about treatment effects."

From the perspective of treatment choice, the Campbell-Rosenbaum position is well grounded if treatment response is homogeneous. Then researchers can aim to learn about treatment response in easy-to-analyze study populations and planners can be confident that research findings can be extrapolated to populations of interest. In human populations, however, homogeneity of treatment response may be the exception rather than the rule. Whether the context be medical or educational or social, it is common to find that people vary in their response to treatment. To the degree that treatment response is heterogeneous, a planner cannot readily extrapolate research findings from a study population to a population of interest, as optimal treatments in the two may differ. Hence correspondence between the study population and the population to be treated assumes considerable importance. In particular, the study population and the population of interest should have the same mean treatment response $\{E[y(\cdot)|x)], x \in X\}$ conditional on the observed covariates x.

A specific instance of the general issue arises in research on partial compliance in randomized experiments. Suppose that experimental subjects are drawn at random from the population to be treated, but some do not comply with their designated treatments. The literature offers various prescriptions for analysis of treatment response without imposing assumptions on the compliance process. One is to report findings on intention-to-treat, which keeps attention focused on the population of interest while assuming that compliance in actual treatment conditions will be the same as in the experiment. Another is to report bounds on classical treatment effects. This again keeps attention focused on the population of interest but differs from intention-to-treat analysis as it assumes full compliance in actual treatment conditions; see Section 3.3.

A third proposal is to report treatment effects for the sub-population of "compliers," persons who would comply with their designated experimental treatments whatever they might be (Imbens and Angrist, 1994; Angrist, Imbens, and Rubin, 1996). This approach assumes full compliance in actual treatment conditions but differs from the bounds analysis in that it makes the study population (i.e., compliers) a particular subset of the population of interest. A planner can extrapolate findings on treatment effects for compliers to the population of interest if treatment response is homogeneous but not to the degree that it is heterogeneous. Indeed, a planner cannot even apply findings for compliers to make treatment choices in this particular subpopulation. The reason is that compliers are not individually identifiable. Each subject in an experiment is placed in one of a set of mutually exclusive treatment groups; hence it is not possible to observe whether a given person would comply with all possible treatment designations.

From the perspective of treatment choice in heterogeneous populations, I see no reason to give internal validity primacy relative to external validity. I am unable to motivate interest in the sub-population of compliers. To be fair, researchers who have stressed internal validity and those who have recently focused attention on compliers have not necessarily asserted that the objective of their research is to inform treatment choice. For example, Angrist, Imbens, and Rubin (1996) view their goal as the discovery of "causal effects," without reference to a treatment-choice problem.

3. Identification Problems and Treatment Choice Under Ambiguity

Ideally, a planner facing the treatment choice problem described in Section 2 would like studies of treatment response to reveal in full how mean outcomes vary with treatments and covariates. In practice, problems of identification and statistical inference limit the information that studies can provide. Statistical and identification problems are logically distinct, so it is analytically useful to consider them in isolation from one another. In this section on identification, I suppose that researchers are able to draw random samples of unlimited size from their study populations and hence know (almost surely) whatever population features their sampling processes are capable of revealing. In Section 4, I suppose that researchers are only able to draw random samples of finite size and hence must make statistical inferences about their study populations.

Whereas planners reading the research literature routinely find measures of statistical imprecision, they may obtain little sense of how identification problems limit inference on treatment response. The prevalent practice among researchers has been to report point estimates of treatment effects obtained by combining the available data with *structural assumptions* – suppositions about treatment selection and response – strong enough to yield point identification. However the estimates so obtained often have rather fragile foundations, as becomes plain from observing the persistent disagreements among researchers about the plausibility of different identifying assumptions.

I have long argued that researchers and planners alike would be better served if the customary practice were to first report the inferences that are possible using only knowledge of the sampling process generating the data, then report inferences under weak but highly credible structural assumptions, and finally report traditional point estimates under strong assumptions (e.g., Manski, 1989, 1995; Manski and Nagin, 1998; Horowitz and Manski, 1998, 2000). Readers of research literature would then be able to draw information from study findings whether or not they accept all of the assumptions used to generate point estimates.

With this in mind, Section 3.1 continues the formalization of the planner's problem begun in Section 2.1. Sections 3.2 and 3.3 then consider the implications for treatment choice of the selection problem, perhaps the most fundamental identification problem arising in studies of treatment response. Section 3.2 addresses the selection problem in observational studies, while Section 3.3 examines its manifestation in randomized experiments with partial compliance. Section 3.4 describes the identification problem that arises in studies with missing outcome and covariate data, and illustrates with data from a clinical trial of treatments for hypertension.

3.1. The Planner's Problem, Continued

By equations (3) and (4), the planner would like to choose a treatment rule that solves the optimization problem

$$(7)\quad \max_{z(\cdot) \in Z} \int \sum_{t \in T} E[y(t)|x]\cdot 1[z(x) = t]\, dP(x).$$

The covariates x are observable, so it is realistic to assume that the planner can learn the distribution P(x) of covariates in the population to be treated. Research on treatment response is motivated by the planner's desire to learn $\{E[y(\cdot)|x)], x \in X\}$.

Identification problems limit the information that studies of treatment response provide. Considering the matter in abstraction, suppose that a planner learns from the available studies that mean treatment response conditional on the observed covariates lies in some set H; that is, $\{E[y(\cdot)|x], x \in X\} \in H$, for some $H \subset Y^T \times X$. This information may not suffice to solve problem (7), in which case the planner faces a problem of treatment choice under *ambiguity*. What should the planner do in such a situation?

Clearly the planner should not choose a dominated treatment rule. A feasible treatment rule $z(\cdot)$ is dominated if there exists another feasible rule, say $z'(\cdot)$, which necessarily yields at least the social welfare of $z(\cdot)$ and which performs strictly better than $z(\cdot)$ in some possible state of nature. That is, $z(\cdot) \in Z$ is dominated if there exists a $z'(\cdot) \in Z$ such that

$$(8a)\quad \int \sum_{t \in T} \eta(t, x)\cdot 1[z(x) = t]\, dP(x) \;\le\; \int \sum_{t \in T} \eta(t, x)\cdot 1[z'(x) = t]\, dP(x)\, , \quad \forall\, \eta \in H$$

and

$$(8b)\quad \int \sum_{t \in T} \eta(t, x)\cdot 1[z(x) = t]\, dP(x) \;<\; \int \sum_{t \in T} \eta(t, x)\cdot 1[z'(x) = t]\, dP(x)\,, \quad \text{some } \eta \in H,$$

where $\eta(t, x)$ denotes a feasible value of $E[y(t) \mid x]$.

The central difficulty of treatment choice under ambiguity is that there is no clearly best way for the planner to choose among treatment rules that are undominated. The most that can be said is that the literature on decision theory suggests a variety of "reasonable" ways to choose among undominated alternatives. Two particularly common suggestions are use of the *maximin rule* or of a *Bayes decision rule*.

Wald (1950) proposed that a person facing a problem of choice under ambiguity select an action that maximizes the minimum welfare attainable under all possible states of nature. In the present treatment choice setting, this means solution of the optimization problem

$$(9)\quad \max_{z(\cdot) \in Z^*} \; \min_{\eta \in H} \; \{\int \sum_{t \in T} \eta(t, x)\cdot 1[z(x) = t]\, dP(x)\},$$

where Z^* denotes the set of undominated treatment rules. Bayesian decision theorists recommend that a person facing ambiguity place a subjective distribution on the states of nature and maximize expected welfare with respect to this distribution. In the treatment choice context, the planner would place a σ-algebra Σ and a probability measure π on the set H, where π expresses the decision maker's personal beliefs about where $\{E[y(\cdot) \mid x], x \in X\}$ may lie within H. The planner would then solve the optimization problem

$$(10)\quad \max_{z(\cdot) \in Z^*} \; \int \{\int \sum_{t \in T} \eta(t, x)\cdot 1[z(x) = t]\, dP(x)\}\, d\pi.$$

Although the maximin rule and Bayes rules differ in their details, they share a key common feature. In both cases the planner replaces the original optimization problem (7), which is not solvable, with another optimization problem that is solvable, namely (9) or (10). See Manski (2000) for further discussion.

3.2. The Selection Problem in Observational Studies

The particular form of the set H expressing the planner's identification problem depends on the available studies of treatment response. A familiar scenario envisions a planner who seeks to draw information from an observational study of a population that corresponds to the population of interest, in the sense of Section 2.3. It is well known that empirical inference on treatment response faces a fundamental difficulty, the selection problem, when the process of treatment selection in the study population is not known. However the precise nature of the selection problem has not been as well appreciated, and it is important that it be understood.

Throughout this section, the population to be treated is denoted J, as earlier. The study population is henceforth denoted J_0. Treatments are yet to be assigned to the members of J, but treatments in J_0 have previously been assigned and outcomes realized. To simplify the exposition, I assume that populations J and J_0 correspond in the strong sense that they have the same distribution $P[y(\cdot), x]$ of response functions and covariates. With some increase in notational burden, the analysis extends immediately to cases in which J and J_0 only share the same mean treatment responses $\{E[y(\cdot)|x], x \in X\}$. If the study population and the population of interest do not correspond in at least this manner, a planner wanting to use an observational study to guide treatment choice must cope not only with the selection problem examined here but also with the extrapolation problem discussed in Section 2.3.

Let $s(\cdot): J_0 \rightarrow T$ denote the *status quo* treatment rule; that is, the rule actually applied in the study population. Then an observational study of population J_0 can reveal the realized (covariate, treatment, outcome) triples $\{x_j, s(j), y_j[s(j)]; j \in J_0\}$. Under the maintained assumption that populations J and J_0 are distributionally identical, observation of J_0 reveals the distribution P[x, s, y(s)] of (covariate, treatment, outcome) triples that would be realized in population J if treatment rule s(·) were to be applied there. The question is: What does knowledge of P[x, s, y(s)] reveal about mean treatment response $\{E[y(\cdot)|x], x \in X\}$?

3.2.1. No-Assumptions Bounds on Mean Treatment Response

Let K_0 and K_1 denote the lower and upper endpoints of the logical range of the response functions. If outcomes are binary, for example, then $K_0 = 0$ and $K_1 = 1$. If outcomes are unbounded, then $K_0 = -\infty$ and $K_1 = \infty$. For each $t \in T$ and $x \in X$, use the law of iterated expectations to write

$$(11)\quad E[y(t)|x] \;=\; E[y(t)|x, s = t]\cdot P(s = t|x) + E[y(t)|x, s \neq t]\cdot P(s \neq t|x).$$

Empirical knowledge of P[x, s, y(s)] implies knowledge of $E[y(t)|x, s = t]$, $P(s = t|x)$, and $P(s \neq t|x)$ but reveals nothing about $E[y(t)|x, s \neq t]$. We know only that the last quantity lies in the interval $[K_0, K_1]$. Hence $E[y(t)|x]$ lies within this sharp bound (Manski, 1989, 1990):

$$(12)\quad E[y(t)|x, s = t]\cdot P(s = t|x) + K_0\cdot P(s \neq t|x) \;\leq\; E[y(t)|x] \;\leq\; E[y(t)|x, s = t]\cdot P(s = t|x) + K_1\cdot P(s \neq t|x).$$

It follows that in an observational study maintaining no structural assumptions, the set H is the $(Y^T \times X)$-dimensional rectangle

$$(13)\quad \mathrm{H} \;=\; [E[y(t)|x, s = t]\cdot P(s = t|x) + K_0\cdot P(s \neq t|x),\; E[y(t)|x, s = t]\cdot P(s = t|x) + K_1\cdot P(s \neq t|x)],$$
$$t \in T,\ x \in X.$$

Inspection of (13) shows that H is a bounded, proper subset of $(Y^T \times X)$ if the response function y(·) has bounded range. I shall henceforth suppose that this basic condition is met. Then we may, without loss of generality, rescale the outcomes y to lie in the unit interval. Setting $K_0 = 0$ and $K_1 = 1$ gives H the particularly simple form

$$(14)\quad H = [E[y(t)\,|\,x, s = t]\cdot P(s = t\,|\,x),\ E[y(t)\,|\,x, s = t]\cdot P(s = t\,|\,x) + P(s \neq t\,|\,x)],\quad t \in T, x \in X.$$

3.2.2. All Treatments are Undominated

The above shows that when outcomes are bounded, an observational study conveys information about mean treatment response even if no structural assumptions are maintained. But is this information enough for the planner to eliminate some treatment rules as dominated? Examination of the structure of the set H shows that this question has an unpleasant answer: All feasible treatment rules are undominated.

To see this, compare two treatment rules. Under one rule, all persons with covariates x receive treatment t'. Under the other rule, all such persons receive a different treatment, say t". In the absence of any empirical evidence on treatment response, we would be able to say only that $E[y(t'')\,|\,x] - E[y(t')\,|\,x]$ lies in the interval [-1, 1]. With the available empirical evidence, (14) yields a narrower bound on $E[y(t'')\,|\,x] - E[y(t')\,|\,x]$. The sharp lower (upper) bound is the lower (upper) bound on $E[y(t'')\,|\,x]$ minus the upper (lower) bound on $E[y(t')\,|\,x]$. Thus

$$\begin{aligned}(15)\quad & E[y(t'')\,|\,x, s = t'']\cdot P(s = t''\,|\,x) - E[y(t')\,|\,x, s = t']\cdot P(s = t'\,|\,x) - P(s \neq t'\,|\,x) \\ & \leq E[y(t'')\,|\,x] - E[y(t')\,|\,x] \\ & \leq E[y(t'')\,|\,x, s = t'']\cdot P(s = t''\,|\,x) + P(s \neq t''\,|\,x) - E[y(t')\,|\,x, s = t']\cdot P(s = t'\,|\,x).\end{aligned}$$

This bound is a subset of the interval [-1, 1]. Its width is $P(s \neq t''\,|\,x) + P(s \neq t'\,|\,x)$, which can be no smaller than one. Hence the bound (15) necessarily contains the value zero. Thus the empirical evidence alone does not reveal which treatment, t' or t'', yields the larger mean outcome. The same reasoning holds for all pairs of treatments and for all values of x. Hence all feasible treatment rules are undominated.

3.2.3. Treatment Choice Using the Maximin Rule and the Ignorable Selection Rule

The fact that all treatment rules are undominated does not imply that the planner should be paralyzed, unwilling and unable to choose a treatment rule. In particular, the planner might apply the conservative maximin rule. This calls for each person with covariates x to receive the treatment that maximizes the lower bound in (12). Thus treatment rule $z^*(\cdot)$ is maximin if , for each $x \in X$, $z^*(x)$ solves the problem

$$(16) \quad \max_{t \in T} E[y(t) \mid x, s = t] \cdot P(s = t \mid x).$$

Application of the maximin rule guarantees that the social welfare achieved by the planner is no less than $E\{\max_{t \in T} E[y(t) \mid x, s = t] \cdot P(s = t \mid x)\}$.

The maximin rule (16) is simple to apply and to comprehend. From the maximin perspective, the desirability of each treatment increases with (i) the mean outcome realized by persons in the study population who received the treatment and (ii) the fraction of persons who received the treatment. The second factor gives form to the conservatism of the maximin rule – the more prevalent a treatment was in the study population, the more expedient it is to choose this treatment in the population of interest.

Observe that $E[y(t) \mid x, s = t] \cdot P(s = t \mid x) \;=\; E\{y(t) \cdot 1[s = t] \mid x\}$. So an alternative way to write the maximin rule is

$$(16') \quad \max_{t \in T} E\{y(t) \cdot 1[s = t] \mid x\}.$$

Thus a planner who uses the maximin rule may be thought of as replacing the outcome of interest y(t) with the composite outcome $y(t)\cdot 1[s = t]$, whose value is always observed. This done, the planner chooses the optimal treatment rule as in (5). Moreover, the planner values covariate information as described in (6).

It is interesting to compare the maximin rule with the one that results if the planner makes the strong identifying assumption that treatment selection is ignorable conditional on x. This familiar structural assumption implies that $E[y(t)\,|\,x] = E[y(t)\,|\,x, s = t]$, so the original optimization problem (7) becomes

$$(17)\quad \max_{t \in T} E[y(t)\,|\,x, s = t].$$

Comparison of (16) and (17) shows how the assumption of ignorable treatment selection supplants the conservatism of the maximin rule; the fraction of the study population who received a given treatment no longer is germane. Observe that the maximin rule and the ignorable selection rule yield the same treatment choice for persons with covariates x if the vectors $\{E[y(t)\,|\,x, s = t], t \in T\}$ and $\{P(s = t\,|\,x), t \in T\}$ have the same rank order. Otherwise, the two rules may yield different treatment choices.

A planner who uses the data from an observational study to implement the ignorable selection rule can defend the resulting treatment choices as consistent with the available empirical evidence. Inspection of (14) shows that the vector $\{E[y(t)\,|\,x, s = t], t \in T, x \in X\}$ lies in the set H; hence an observational study does not reveal enough about treatment response to enable rejection of the hypothesis that treatment selection is ignorable. Of course, consistency of a treatment rule with the available evidence does not imply optimality of the rule. Given only the information available from an observational study, a planner can equally well defend picking out any element of the set H and choosing treatments as if this is the actual value of $\{E[y(t)\,|\,x], t \in T, x \in X\}$.

3.2.4. Instrumental Variable Assumptions

The above discussion considers the situation of a planner who seeks to use the empirical evidence of an observational study, but brings to bear no prior knowledge of treatment selection and response in the study population. A planner might combine the available data with structural assumptions that identify mean response functions and then solve the optimization problem (7). The ignorable selection rule (17) is just one of many possibilities. However, structural assumptions strong enough to identify mean response are difficult to motivate in practice. Imposing assumptions that are not credible does not really eliminate ambiguity in treatment choice.

Between the poles of no structural assumptions and of assumptions that identify mean treatment response lies a vast middle ground of assumptions that are weak enough to be credible but strong enough to shrink the set H relative to its form in equation (14). A literature investigating such assumptions has begun to take shape in the past ten years, with particular emphasis on *instrumental variables* (IV) assumptions. Other middle-ground assumptions that will not be discussed here are examined in Manski (1995, 1997) and Manski and Nagin (1998).

Manski (1990) determined the identifying power of the traditional IV assumption of econometrics, which holds that mean response is constant across sub-populations defined by different values of some covariate. Let $x \equiv (w, v)$. Covariate v, taking values in a space V, is said to be an instrumental variable (in the sense of mean independence) if, for $t \in T$, each value of w, and all $(u, u') \in (V \times V)$,

$$(18)\quad E[y(t)|w, v = u'] \;=\; E[y(t)|w, v = u].$$

When (18) holds, the common value of $E[y(t)|w, v = u]$, $u \in V$ must lie in the intersection of the no-assumptions bounds (12) across the elements of V. Any point in this intersection is feasible. Thus, for all $u \in V$, we obtain the sharp IV bound

$$(19)\quad \sup_{u' \in V} [E(y|w, v = u', s = t) \cdot P(s = t|w, v = u') + K_0 \cdot P(s \neq t|w, v = u')]$$

$$\leq E[y(t)|w, v = u] \leq$$

$$\inf_{u' \in V} [E(y|w, v = u', s = t) \cdot P(s = t|w, v = u') + K_1 \cdot P(s \neq t|w, v = u')].$$

Inspection of (19) shows that an IV assumption has identifying power if the outcomes y have bounded range and if the no-assumptions bounds for $E[y(t)|w, v = u]$, $u \in V$ do not coincide.

Subsequently, Hotz, Mullins, and Sanders (1997) and Manski and Pepper (2000) have studied weaker forms of assumption (18). Hotz, Mullin, and Sanders (1997) investigate *contaminated instrument* assumptions; these suppose that a mean-independence assumption holds in the population of interest, but the study population is a probability mixture of the population of interest and one in which the assumption does not hold. Manski and Pepper (2000) determine the identifying power of *monotone instrumental variable* assumptions, which replace the equality in (18) with a weak inequality.

Other authors have studied stronger forms of IV assumptions. Robins (1989) and Balke and Pearl (1997), motivated by the problem of inference in randomized experiments with partial compliance, begin from the statistical independence assumption

$$(20)\quad P[y(\cdot)|w, v = u'] = P[y(\cdot)|w, v = u],$$

which implies the mean independence assumption (18). Robins (1989) used (20) to derive the same bound (19) as independently reported in Manski (1990) under assumption (18). Later Balke and Pearl (1997), focusing on situations in which outcomes are binary, showed that this bound is not always sharp when statistical independence holds; the sharp bound solves a particular linear programming problem. A different strengthening of (18) is investigated by Heckman and Vytlacil (2000), who pose a nonparametric latent-variable model of treatment selection. Their model yield a sharp bound on mean treatment response whose form is a special case of the IV bound (19).

Whichever form of IV assumption one may impose, the generic result is that the assumption shrinks the set H relative to its no-assumption form (14). IV assumptions typically do not reduce H to a point, but they often do shrink H enough for a planner to conclude that some treatment rules are dominated. Even when this is not the case, a planner may find such an assumption useful. For example, it may enable implementation of a refined version of the maximin rule discussed in Section 3.2.3.

It is important to distinguish the present use of IV assumptions to inform treatment choice from their recent application by Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) to identify *local average treatment effects* (LATE). An LATE is an average treatment effect within the non-observable sub-population of the study population who would have selected a different treatment had they realized a different value of the instrumental variable. (In the context of randomized experiments, these are the "compliers" discussed in Section 2.3.) If treatment response is homogeneous, the LATE is the traditional IV estimate of treatment effects (e.g., Heckman and Robb, 1985). If treatment response is heterogeneous, a planner cannot use an LATE to infer treatment effects in the population of interest and, indeed, cannot even identify the persons who form the sub-population to which an LATE applies.

3.3. The Selection Problem in Randomized Experiments with Partial Compliance

As earlier, let J denote the population of interest and let J_0 denote a study population. A randomized experiment establishes multiple treatment groups within J_0, each with a different designated treatment. Randomization implies that, abstracting from finite-sampling variation, each treatment group has the same distribution of response functions and covariates as does the study population J_0 itself. Let J_m, $m \in M \subset T$ denote the treatment groups; thus the study designer intends that persons in treatment group J_m receive treatment m. Let $s_m(\cdot)\colon J_m \rightarrow T$ denote the treatment rule actually prevailing in group J_m.

From the perspective of a planner choosing a treatment rule, an ideal experiment satisfies two conditions. First, the study population J_0 should correspond to the population of interest J, having the same distribution of response functions and covariates, or at least the same mean treatment responses. Second, the members of each treatment group should comply with their designated treatments; for each $m \in M$, rule $s_m(\cdot)$ should have the form $s_m(j) = m$, $j \in J_m$. Satisfaction of these conditions implies that the experiment reveals the mean treatment responses $\{E[y(m)|x], m \in M, x \in X\}$ that the planner wishes to learn.

The present discussion assumes that the first condition holds and considers the selection problem that arises when the second condition does not hold. Partial compliance with designated treatments is a close to ubiquitous occurrence in randomized experiments; hence planners need to understand the implications. Moreover, it is enlightening to juxtapose the present selection problem with its form in observational studies.

Section 3.3.1 examines the simple, common case in which experimental subjects who do not comply with their designated treatments cannot "cross over" to receive other experimental treatments; instead they receive some "usual" treatment available outside the experimental setting. Section 3.3.2 considers experiments with crossover.

3.3.1. Experiments without Crossover

A common objective of randomized experiments is to learn the properties of new treatments that planners have not previously been available. Let $M \subset T$ denote these new treatments and let the remaining treatments T - M denote "usual" treatments that are available outside the experimental setting. It is often realistic to suppose that, through past experience, planners know the mean outcomes $\{E[y(t)|x], t \in T - M, x \in X\}$ of the usual treatments. The problem is to learn about response to the new treatments.

In experiments of this type, a subject who does not comply with his or her designated treatment is not permitted to cross over to another experimental treatment. Such a subject may, however, receive one of the usual treatments T - M. Hence the only persons in the study population who receive a new treatment m

$\in$ M are the experimental subjects who are placed in treatment group J_m and who comply with their designated treatment.

As in Section 3.2, suppose that the outcomes y are bounded and are scaled to lie in the unit interval. Then, as earlier, the law of iterated expectations

$$(21)\quad E[y(m)|x] = E[y(m)|x,s_m = m]\cdot P(s_m = m|x) + E[y(m)|x, s_m \neq m]\cdot P(s_m \neq m|x)$$

implies that E[y(m)|x] lies in the interval

$$(22)\quad E[y(m)|x, s_m = t]\cdot P(s_m = m|x) \leq E[y(m)|x] \leq E[y(m)|x, s_m = m]\cdot P(s_m = m|x) + P(s_m \neq m|x).$$

Hence the set H of feasible values for {E[y(t)|x], t $\in$ T, x $\in$ X} is the $(Y^T \times X)$-dimensional rectangle

$$(23)\quad H = \{[E[y(m)|x, s_m = m]\cdot P(s_m = m|x),\ E[y(m)|x, s_m = m]\cdot P(s_m = m|x) + P(s_m \neq m|x)],\ m \in M\},$$
$$\{E[y(t)|x], t \in T- M\}, \qquad x \in X.$$

The set H here bears some similarity to that obtained in equation (14) when the data are generated by an observational study. However, the present form of H does not carry the earlier implication that all feasible treatment rules are undominated. This is easiest to see when, for persons with covariates x, we compare a rule assigning a new treatment m with another rule assigning a usual treatment t. Then (23) implies this sharp bound on E[y(m)|x] - E[y(t)|x]:

$$(24)\quad E[y(m)|x, s_m = m]\cdot P(s_m = m|x) - E[y(t)|x] \leq E[y(m)|x] - E[y(t)|x]$$
$$\leq E[y(m)|x, s_m = m]\cdot P(s_m = m|x) + P(s_m \neq m|x) - E[y(t)|x].$$

Treatment m is necessarily preferred to t if the lower bound in (24) is greater than zero, and t must be preferred to m if the upper bound is less than zero.

The planner may also be able to rank order new treatments relative to one another. Let m' and m'' be two such treatments. The sharp bound on $E[y(m'')|x] - E[y(m')|x]$ is

$$(25)\quad E[y(m'')|x, s_{m''} = m'']\cdot P(s_{m''} = m''|x) - E[y(m')|x, s_{m'} = m']\cdot P(s_{m'} = m'|x) - P(s_{m'} \neq m'|x)$$
$$\leq E[y(m'')|x] - E[y(m')|x]$$
$$\leq E[y(m'')|x, s_{m''} = m'']\cdot P(s_{m''} = m''|x) + P(s_{m''} \neq m''|x) - E[y(m')|x, s_{m'} = m']\cdot P(s_{m'} = m'|x).$$

This bound has width $P(s_{m''} \neq m''|x) + P(s_{m'} \neq m'|x)$, which is less than one except in severe cases of noncompliance. When its width is less than one, the bound (25) may lie entirely to one side of zero, depending on the specifics of the case.

In other respects, treatment choice using data from an experiment without crossover is analogous to treatment choice using observational data. A planner lacking credible structural assumptions may use the maximin rule, the ignorable selection rule, or other rules consistent with the empirical evidence to choose among the treatments that are undominated. A planner who can impose credible assumptions, perhaps IV assumptions as discussed in Section 3.2.4, may be able to shrink the set H relative to its form in (23) and make better treatment choices.

3.3.2. Experiments with Crossover

Experiments without crossover are simple to analyze because only one part of the study population receives each new treatment m, these being the experimental subjects who are placed in treatment group J_m and who comply with their designated treatment. Experiments with crossover are more complex, but also more informative.

As earlier, suppose that the outcomes y are bounded and are scaled to lie in the unit interval. Applying the law of iterated expectations

$$(26)\quad E[y(m)|x] \;=\; E[y(m)|x,s_n = m]\cdot P(s_n = m|x) + E[y(m)|x,\, s_n \neq m]\cdot P(s_n \neq m|x)$$

to all of the treatment groups $n \in M$ shows that $E[y(m)|x]$ lies in all of the intervals

$$(27)\quad E[y(m)|x,\, s = t]\cdot P(s_n = m|x) \;\leq\; E[y(m)|x] \;\leq\; E[y(m)|x,\, s_n = m]\cdot P(s_n = m|x) + P(s_n \neq m|x).$$

Hence $E[y(m)|x]$ lies in the intersection of these intervals, namely

$$(28)\quad \sup_{n \in M} \{E[y(m)|x,\, s_n = t]\cdot P(s_n = m|x)\} \;\leq\; E[y(m)|x] \;\leq \inf_{n \in M} \{E[y(m)|x,\, s_n = m]\cdot P(s_n = m|x) + P(s_n \neq m|x)\}.$$

This bound has the same form as the IV bound (19), derived by assuming the existence of an instrumental variable in the sense of mean independence. Indeed, randomization of designated treatments implies that the treatment group within which a subject is placed is such an IV. Thus (28) is formally a special case of the IV bound (19). An experiment without crossover yields a special case of (28), in which $P(s_n = m|x) \;=\; 0$ when $n \neq m$.

The above suffices to show that an experiment with crossover is more informative than one without crossover. However, the bound (28) need not be sharp. The reason is that randomization implies more than that a subject's designated treatment is an IV in the sense of mean independence; it is an IV in the stronger sense of statistical independence given in (20). Characterization of the sharp bound on $\{E[y(m)|x], m \in M\}$ implied by an experiment with crossover appears to be a quite complex problem. As noted earlier, Balke and

Pearl (1997) have examined the special case in which outcomes are binary and shown that the sharp bound solves a particular linear programming problem. They find that the sharp bound improves on (28) for some data configurations but not for others. Unfortunately, a succinct interpretation of their finding has been elusive. Moreover, the form of the sharp bound when outcomes are not binary remains entirely an open question. This being the case, Robins and Greenland (1996) suggest use of the relatively simple albeit not necessarily sharp bound (28) as a practical expedient.

3.4. Treatment Choice Using Studies with Missing Covariate and Outcome Data

The selection problem manifests itself as the absence of data on outcomes that would have been realized by members of the study population had they received other treatments. Studies of treatment response commonly have missing data for other reasons as well. Researchers performing randomized experiments may encounter data collection problems at the beginning of a trial that result in missing covariate data for some subjects. Subsequently, noncompliance and/or attrition of subjects from the trial may prevent collection of complete outcome data. Similar missing data problems occur in observational studies, where covariate data may be missing due to survey nonresponse and outcome data may be missing due to the selection problem and/or survey nonresponse.

Section 3.4.1 summarizes recent research on the identification problem created by missing data. Section 3.4.2 uses a particular study to draw out the implications for treatment choice.

3.4.1. Identification of Mean Treatment Response in Studies with Missing Data

Researchers have typically coped with the identification problem created by missing data by imposing structural assumptions strong enough to identify mean treatment response. It has been particularly common to assume that covariate and outcome data are missing at random. Occasionally, a model of nonrandom missing data is asserted. Either way, the identification problem is resolved and precision of statistical inference becomes the central concern. See, for example, Little (1992), Robins, Rotnitzky, and Zhao (1994), and Wang, Wang, Zhao, and Ou (1997).

Recent research by the author and Joel Horowitz addresses the problem of inference on mean treatment response without imposing assumptions on the distribution of missing data (Horowitz and Manski, 1998, 2000). In principle, the set H of feasible values of $\{E[y(\cdot)|x], x \in X\}$ can be constructed by entertaining all possible distributions of the missing data. The practical problem is to characterize H in a useful manner. An important part of this problem is solved in Theorem 1 of Horowitz and Manski (2000), which supposes that outcomes are binary and derives a closed-form, albeit somewhat intricate, bound on $E[y(\cdot)|x]$. The theorem applies to general missing data problems; some observations may be complete, some may have missing outcome data, others may have missing covariate data, and still others may have jointly missing covariate and outcome data. The bound is sharp in generic observational studies and in experiments without crossover. It also holds in experiments with crossover, but may not be sharp in such cases.

Beyond the provision of specific identification findings of practical use, my recent research on missing data problems reveals the general truth that missing covariate data and missing outcome data have quite different implications for identification. Observations with missing outcome data, or with jointly missing outcome and covariate data, are uninformative about the distribution of treatment response. However, observations in which only covariate data are missing do carry information about treatment response. The consequences are made plain in Horowitz and Manski (1998), which considers general real-valued outcomes and examines three polar observational cases: only outcome data are missing, only covariate data are missing, covariate and outcome data are jointly missing. Analysis of the selection problem showed that a simple bound on mean treatment response results when only outcomes are missing. We find that jointly missing covariate and outcome data imply a similarly simple bound; its form is the same as when only outcome data are missing, except that the *effective* rate of missing data is higher than the actual rate. In contrast, the case in which only covariate data are missing yields a different bound that solves a complex extremum problem. When outcomes are continuous, this extremum problem must be solved numerically. When outcomes are binary, the bound has a tractable closed-form expression.

3.4.2. Choosing Treatments for Hypertension Using Data from a Trial with Missing Data

Physicians routinely face the problem of choosing treatments for hypertension. Medical research has sought to provide guidance through the conduct of clinical trials comparing alternative treatments. Such trials inevitably have missing data. I illustrate here how physicians might use the data from a recent trial to inform treatment choice, without imposing assumptions about the distribution of the missing data.

Materson *et al.* (1993) and Materson, Reda, and Cushman (1995) present findings from a U.S. Department of Veteran Affairs (DVA) clinical trial of treatments for hypertension. Male veteran patients at 15 DVA hospitals were randomly assigned to one of 6 antihypertensive drug treatments or to placebo: hydrochlorothiazide ($t = 1$), atenolol ($t = 2$), captopril ($t = 3$), clonidine ($t = 4$), diltiazem ($t = 5$), prazosin ($t = 6$)}, placebo ($t = 7$). The trial had two phases. In the first, the dosage that brought diastolic blood pressure (DBP) below 90 mm Hg was determined. In the second, it was determined whether DBP could be kept below 95 mm Hg for a long time. Treatment was defined to be successful if DBP < 90 mm Hg on two consecutive measurement occasions in the first phase and DBP ≤ 95 mm Hg in the second. Treatment is unsuccessful otherwise. Thus the outcome of interest is binary, with $y = 1$if the criterion for success is met and $y = 0$ otherwise. Materson *et al.* (1993) recommend that physicians making treatment choices should consider this medical outcome variable as well as patient's quality of life and the cost of treatment.

Among various covariates measured at the time of randomization, one was the biochemical indicator "renin response," taking the values $x = (\text{low, medium, high})$. This covariate has previously been studied as a factor that might be related to the probability of successful treatment (Freis, Materson, and Flamenbaum 1983). Renin-response data were missing for some patients in the trial. Moreover, some patients dropped out of the trial before their outcomes could be determined. Crossover from one treatment group to another was not permitted. The pattern of missing covariate and outcome data is shown in Table 1 of Horowitz and Manski (2000), reproduced here.

Table 1: Missing Data in the DVA Hypertension Trial

Treatment	Number Randomized	Observed Successes	None Missing	Missing Only y	Missing Only x	Missing y and x
1	188	100	173	4	11	0
2	178	106	158	11	9	0
3	188	96	169	6	13	0
4	178	110	159	5	13	1
5	185	130	164	6	14	1
6	188	97	164	12	10	2
7	187	57	178	3	6	0

Horowitz and Manski (2000) applied our Theorem 1 to the DVA data to estimate sharp bounds on the success probabilities $\{P[y(t) = 1 | x], t = 1, ..., 7\}$ for each value of x, without imposing assumptions on the distribution of missing data. Rather than report the bounds on the success probabilities directly, we reported the implied bounds on the average treatment effects $\{P[y(t) = 1 | x] - \{P[y(7) = 1 | x], t = 1, ..., 6\}$, which measure the efficacy of each treatment relative to the placebo. This reporting decision was motivated by the traditional research problem of testing the hypothesis of zero treatment effect. We did not explicitly examine the implications for treatment choice.

Table 2 reports the estimates of the bounds on the success probabilities themselves. To keep attention focused on the identification problem, suppose that the estimates are the actual bounds rather than finite sample estimates. Consider a physician who accepts the DVA success criterion, observes renin response, and has no prior information on mean treatment response or the distribution of missing data. How might this physician choose treatments in a population analogous to that studied in the DVA trial?

Table 2: No-Assumptions Bounds on Success Probabilities Conditional on Renin Response

Renin				Treatment			
Reponse	1	2	3	4	5	6	7
Low	[.54, .61]	[.52, .62]	[.43, .53]	[.58, .66]	[.66, .76]	[.54, .65]	[.29, .32]
Medium	[.47, .62]	[.60, .74]	[.53, .68]	[.50, .69]	[.68, .85]	[.41, .65]	[.27, .32]
High	[.28, .50]	[.64, .86]	[.56, .75]	[.63, .84]	[.55, .78]	[.34, .59]	[.28, .40]

First, the physician should eliminate from consideration the dominated treatments, indicated by shading in the table. For patients with low renin response, treatments 1, 2, 3, 4, 6, and 7 are all dominated by treatment 5, which has the greatest lower bound (.66). For patients with medium renin response, treatments 1, 3, 6, and 7 are dominated by treatment 5, which again has the greatest lowest bound (.68). For patients with high renin response, treatments 1, 6, and 7 are dominated by treatment 2, which has the greatest lowest bound in this case (.64). Thus, without imposing any assumptions on the distribution of missing data, the physician can reject treatments 1, 6, and 7 for all patients, can reject treatment 3 for patients with medium renin response, and can determine that treatment 5 is optimal for patients with low renin response.

In the absence of assumptions about the distribution of missing data, it is not possible to give the physician guidance on how to choose among undominated treatments for patients with medium and high renin response. A physician using the maximin rule would choose treatment 5 for patients with medium renin response and treatment 2 for patients with high renin response. This is a reasonable treatment rule, but one cannot say that it is an optimal rule.

Exploring the reasons for missing data in the VA trial, Horowitz and Manski (2000) do not find a credible basis to impose assumptions on the distribution of missing outcome data, but do find it plausible to entertain the assumption that covariate data are missing completely at random (MCAR). This assumption generates tighter bounds on mean treatment response, given in our Theorem 3. Table 3 presents the resulting estimates of bounds on success probabilities.

Table 3: MCAR-Covariates Bounds on Success Probabilities Conditional on Renin Response

Renin	Treatment						
Reponse	1	2	3	4	5	6	7
Low	[.57, .58]	[.54, .60]	[.44, .49]	[.61, .63]	[.69, .74]	[.56, .62]	[.31, .32]
Medium	[.52, .57]	[.66, .71]	[.59, .59]	[.55, .63]	[.81, .81]	[.46, .57]	[.32, .32]
High	[.35, .35]	[.75, .83]	[.65, .65]	[.77, .77]	[.67, .70]	[.40, .47]	[.33, .40]

These tighter bounds resolve most but not all of the remaining ambiguity in treatment choice. A physician who accepts the assumption that covariate data are MCAR can conclude that treatment 5 is optimal for patients with low and medium renin response. This physician can narrow consideration to treatments 2 and 4 for patients with high renin response, but the data combined with the MCAR assumption do not suffice to choose between these two treatments.

4. Treatment Choice Using Sample Data

To examine the implications for treatment choice of identification problems, it was convenient to suppose that researchers are able to draw random samples of unlimited size from their study populations. In practice, researchers draw samples of finite size. Hence planners seeking to learn from research on treatment response face not only identification problems but also problems of statistical inference. As burdensome as identification problems may be, they at least have the analytical clarity of exercises in deductive logic. Statistical inference is a more murky matter of induction from samples to populations.

Frequentist and Bayesian statistics suggest alternative approaches to decision making using sample data. A thoughtful article by Samaniego and Reneau (1994) makes plain that neither approach dominates the other when the decision problem is to choose a point estimator that minimizes square loss. These authors find that (page 947): “the method to be favored in a particular application depends crucially on the quality of the prior information available, with Bayesian and frequentist methods each emerging as preferable under specific, and complementary, circumstances.” Berger (1985), while generally advocating the Bayesian perspective, makes much the same point when he states (page 121): “A Bayesian analysis may be 'rational' in the weak axiomatic sense, yet be terrible in a practical sense if an inappropriate prior distribution is used.”

It is evident that in treatment choice as in other contexts, there is no uniformly best approach to decision making using sample data. Nevertheless, there is much that researchers can do to help planners. Section 4.1 formalizes the planner’s problem in general terms, defining statistical treatment rules and the expected welfare achieved by a rule. Section 4.2 applies the expected welfare criterion to evaluate the conditional success and unconditional success rules described in Section 1. Section 4.3 draws implications for study design and for the interpretation of research findings. To focus attention on the problem of statistical inference, the discussion in this section assumes that mean treatment response is identified.

4.1. The Planner’s Problem, with Sample Data

4.1.1. Statistical Treatment Rules

Thus far, the feasible treatment rules have been functions that map covariates into treatments. A planner observing sample data can make treatment choices depend on these data. To make this explicit, I now define *statistical treatment rules* to be functions that map covariates and sample data into treatments. The term *statistical treatment rule,* or *STR* for short, recalls Wald (1950), who used the term *statistical decision function* to describe functions that map sample data into decisions.

Considering this in abstraction, let Q denote a sampling process generating the study data and let Ψ denote the associated sample space; that is, Ψ is the set of data samples that may be drawn under Q. Let Z_Ψ denote the space of functions mapping $X \times \Psi$ into T. Then each function $\zeta(\cdot, \cdot) \in Z_\Psi$ defines a statistical treatment rule. The problem of treatment choice using sample data is thus formalized as the problem of evaluating alternative statistical treatment rules. Repeating the reasoning that led previously to the optimization problem (7), the planner now would like to choose an STR that solves the problem

$$(29)\quad \max_{\zeta(\cdot,\,\cdot)\,\in\, Z_\Psi} \int \sum_{t\,\in\, T} E[y(t)\,|\,x]\cdot 1[\zeta(x,\,\psi) = t]\; dP(x).$$

As earlier, I shall assume that the planner knows the covariate distribution P(x) in the population to be treated. The problem is to learn the mean treatment responses $\{E[y(\cdot)\,|\,x],\ x \in X\}$.

At this point the difference between identification and statistical inference becomes apparent. As discussed in Section 3.1, an identification problem may prevent a planner from using a study of treatment response to learn the mean treatment responses, but still enable the planner to place them in some informative set H. Finite-sample data, however, do not suffice to deduce H; they only enable frequentist or Bayesian probabilistic statements. A frequentist can use the data to estimate H and may be able to characterize the sampling distribution of the estimate. A Bayesian can combine subjective prior information with the data to induce a posterior distribution for H. Neither a frequentist nor a Bayesian can address problem (29) as posed.

4.1.2. Frequentist and Bayesian STRs

Unable to deal with problem (29) as posed, frequentists and Bayesians transform the problem into ones that they can address. The standard frequentist prescription is to use the sample data to estimate H and then behave as if the estimate is H. Typically, frequentists impose assumptions strong enough to identify mean treatment response, use the sample data to estimate $\{E[y(t)\,|\,x],\ t \in T,\ x \in X\}$, and then plug the estimate into (29), yielding

$$(30)\quad \max_{\zeta(\cdot,\,\cdot)\,\in\,Z_\Psi} \int \sum_{t\,\in\,T} E_\psi[y(t)\,|\,x]\cdot 1[\zeta(x,\,\psi) = t]\; dP(x),$$

where $E_\psi[y(t)\,|\,x]$ is the estimate of $E[y(t)\,|\,x]$ using the sample data ψ. An STR solves (30) if, for all $x \in X$, it solves the problem $\max_{t\,\in\,T} E_\psi[y(t)\,|\,x]$. The basic argument for selecting such a frequentist STR is asymptotic. If $\{E_\psi[y(\cdot)\,|\,x], x \in X\}$ converges to $\{E[y(\cdot)\,|\,x], x \in X\}$ as sample size increases, then the social welfare yielded by the frequentist rule converges to $E\{\max_{t\,\in\,T} E[y(t)\,|\,x]\}$, the highest welfare achievable.

The standard Bayesian prescription is to place a subjective prior distribution on the population response distribution, use the sample data to induce a posterior distribution on $\{E[y(\cdot)\,|\,x], x \in X\}$ through Bayes Rule, and then maximize expected utility with respect to this posterior. In particular, a risk-neutral Bayesian planner would solve the problem

$$(31)\quad \max_{\zeta(\cdot,\,\cdot)\,\in\,Z_\Psi} \int \sum_{t\,\in\,T} e\{E[y(t)\,|\,x]\,|\,\psi\}\cdot 1[\zeta(x,\,\psi) = t]\; dP(x),$$

where $e\{E[y(t)\,|\,x]\,|\,\psi\}$ is the planner's posterior mean for $E[y(t)\,|\,x]$, given the sample data ψ. An STR solves (31) if, for all $x \in X$, it solves the problem $\max_{t\,\in\,T} e\{E[y(t)\,|\,x]\,|\,\psi\}$. See Dehejia (1999) for a recent application of the Bayesian approach to treatment choice.

The Bayesian prescription is the same whether or not mean treatment response is identified. Nevertheless, identification problems affect a Bayesian planner. If mean treatment response is identified, the posterior $\{e\{E[y(\cdot)\,|\,x]\,|\,\psi\}, x \in X\}$ generally converges to $\{E[y(\cdot)\,|\,x], x \in X\}$ as sample size increases; hence Bayesian STRs have the same desirable asymptotic properties as frequentist ones. If mean response is not identified, however, sample data do not provide the information needed for the posterior to converge; instead, the posterior distribution on $\{E[y(\cdot)\,|\,x], x \in X\}$ inherits properties of the prior.

4.1.3. Evaluation of STRs as Procedures

It is easy enough to prescribe frequentist and Bayesian STRs that a planner might reasonably entertain. It is more challenging to evaluate their performance.

A planner would like to determine how alternative STRs perform when applied to the particular sample data that the planner observes. However, no rule can perform best in every case. Hence it is reasonable to evaluate STRs as *procedures* applied as the sampling process is engaged repeatedly to draw independent data samples. Evaluation of statistical decision functions as procedures has a long history in the literature on statistical decision theory; see Wald (1950), Berger (1985, Section 1.6.2), and Samaniego and Reneau (1994). The usual practice has been to measure the performance of a procedure by the expected value of the welfare it yields in repeated applications. (Formulating the problem as minimization of loss rather than maximization of welfare, Wald calls expected loss in repeated applications the *risk* of a procedure.) I follow this practice here.

Let ζ be any feasible STR. Repeated engagement of the sampling process Q to draw independent samples ψ makes the population mean outcome $\int \sum_{t\in T} E[y(t)|x]\cdot 1[\zeta(x, \psi) = t]\, dP(x)$ a real-valued random variable. Let Q[·] denote the Q-probability of the event in brackets. Then the *expected welfare* yielded by ζ in repeated applications is

$$(32)\quad W(\zeta) \equiv \int [\int \sum_{t\in T} E[y(t)|x]\cdot 1[\zeta(x, \psi) = t]\, dP(x)]\, dQ(\psi) = \int \sum_{t\in T} E[y(t)|x]\cdot Q[\zeta(x, \psi) = t]\, dP(x).$$

Here $Q[\zeta(x, \psi) = t] \equiv \int 1[\zeta(x, \psi) = t]dQ(\psi)$ is the probability of drawing a sample ψ such that rule ζ selects treatment t for persons with covariates x.

The performance of ζ as a procedure may be measured by comparing its expected welfare $W(\zeta)$ with the optimal welfare $E\{\max_{t\in T} E[y(t)|x]\}$ that would be achievable if mean treatment response were known. Thus we may define the *divergence* of ζ from the ideal to be the non-negative quantity

$$(33)\ D(\zeta) \equiv E\{\max_{t \in T} E[y(t)|x]\} - W(\zeta) = \int \{\max_{t \in T} E[y(t)|x] - \sum_{t \in T} E[y(t)|x]\cdot Q[\zeta(x, \psi) = t]\} dP(x).$$

4.2. Evaluation of the CS and US Rules

There are many frequentist and Bayesian STRs that a planner might want to compare using the expected welfare criterion. Among all STRs, perhaps the most familiar is the frequentist conditional success (CS) rule that selects treatments with the best empirical success rates conditional on the observed covariates. The CS rule has clear asymptotic appeal when the data are from classical randomized experiments, in which all subjects comply with their designated treatments and no covariate or outcome data are missing. The rule has the same asymptotic appeal when the data are from observational studies that mimic classical randomized experiments; that is, studies in which there is good reason to assume that treatment selection is ignorable.

The asymptotic appeal of the CS rule does not guarantee that the rule performs well in finite sample settings. A common suggestion when only small samples of persons with particular covariate values are available is to decompose the covariate vector x into two sub-vectors, say $x = (x_1, x_2)$, and choose treatments that maximize empirical success rates conditional only on x_1. The idea is that aggregating outcome data across persons with the same value of x_2 yields larger samples that may enable more precise estimation of mean responses. The extreme form of limited-covariate STRs is the unconditional success (US) rule, which assigns to all members of the population a treatment with the best unconditional empirical success rate.

This section uses the expected welfare criterion to compare the CS and US rules. I assume throughout that the covariate space X is finite, with $P(x) > 0$, all $x \in X$. To keep the exposition simple, I restrict attention to treatment-choice problems with two feasible treatments, $t = 0$ and $t = 1$. I suppose that the data are from a classical randomized experiment with sub-sample sizes N_{xt}, $t \in T$, $x \in X$. Each sub-sample of subjects, denoted N(x, t), realizes outcomes y_j, $j \in N(x, t)$. I assume that mean treatment response conditional on the covariates x is the same in the study population and in the population to be treated. However, the two populations need not have the same covariate distribution.

I evaluate expected welfare under the assumption that the sampling process Q repeats the randomized experiment with the sub-sample sizes $(N_{xt}, t \in T, x \in X)$ held fixed. This is clearly the right sampling process to consider when the data are from an experiment in which the researcher specifies these sub-sample sizes a priori. Randomized experiments are often carried out under an alternative protocol in which the experimenter specifies a priori only the overall treatment group sizes $N_t \equiv \Sigma_{x \in X} N_{xt}, t \in T$. Then the sub-sample sizes $(N_{xt}, t \in T, x \in X)$ are ancillary statistics. Hence conditioning on $(N_{xt}, t \in T, x \in X)$ follows the prescription of statisticians who argue that inference should be performed conditional on ancillary statistics. Others may prefer to evaluate expected welfare under the assumption that the sampling process repeats the experiment with only $(N_t, t \in T)$ held fixed. This requires a more complex analysis than that performed here.

Section 4.2.1 formalizes the CS and US rules, and then examines how expected welfare under these rules depends on sample size and the distribution of treatment response. Section 4.2.2 develops simple conditions on sample size and mean treatment response which imply that the CS rule is superior to the US rule. Section 4.2.3 considers how a planner might choose between the rules in practice, using sample data on treatment response. Section 4.2.4 uses the DVA hypertension trial to illustrate the findings.

The present analysis can be extended to problems with more than two treatments and to comparison of rules that measure success conditional on different sets of covariates. The analysis also applies when the data are from observational studies with ignorable treatment selection. In such studies, the sub-sample sizes $(N_{xt}, t \in T, x \in X)$ are ancillary statistics.

4.2.1. Expected Welfare Under the CS and US Rules

Let $\bar{y}_{xt} \equiv (1/N_{xt}) \sum_{j \in N(x, t)} y_j$ be the sample average outcome among subjects with covariates x assigned to treatment t. Let $\bar{y}_t \equiv \sum_{x \in X} \bar{y}_{xt} \cdot P(x)$ be the population-weighted average outcome among all subjects assigned to treatment t. For each $x \in X$, the CS rule selects a treatment that maximizes $\bar{y}_{xt}$ on $t \in T$. The US rule selects a treatment that maximizes $\bar{y}_t$ on $t \in T$. Each rule requires a tie-breaking convention to be used when multiple treatments maximize the relevant average outcome. I use the convention that treatment 1 is chosen when both treatments yield the same average outcome.

The CS rule is the specific form of the frequentist STR (30) with $E_\psi[y(t)|x] \equiv \bar{y}_{xt}$, a natural nonparametric estimate of $E[y(t)|x]$. The CS rule has clear asymptotic appeal but may not perform well when the sub-samples N_{xt} are small. This motivates interest in the US rule, which aggregates outcome data across values of x into $\bar{y}_t$, a natural estimate of $E[y(t)]$. One might reasonably entertain STRs that use other estimates of $E[y(t)|x]$ and $E[y(t)]$. For example, the literature on Bayesian prediction suggests shrinkage methods that use a weighted average of $\bar{y}_{xt}$ and $\bar{y}_t$ to estimate $E[y(t)|x]$, the weights depending on sample size. A planner possessing prior information restricting $E[y(\cdot)|x]$ might use an estimate that brings this information to bear. Such variations on the CS and US themes warrant attention but will not be examined here.

With the CS and US rules defined as above, the CS rule yields expect welfare

$$(34) \quad W(CS) \;=\; \sum_{x \in X} P(x)\, \{E[y(1)|x] \cdot Q[\bar{y}_{x1} \geq \bar{y}_{x0}] \;+\; E[y(0)|x] \cdot Q[\bar{y}_{x1} < \bar{y}_{x0}]\}.$$

The US rule yields expected welfare

$$(35) \quad W(US) \;=\; E[y(1)] \cdot Q[\bar{y}_1 \geq \bar{y}_0] \;+\; E[y(0)] \cdot Q[\bar{y}_1 < \bar{y}_0].$$

Applying the expected welfare criterion, we shall say that the CS rule is superior or inferior to the US rule if W(CS) is larger or smaller than W(US).

It is easy to see that the CS rule asymptotically yields the optimal population mean outcome. The CS rule is asymptotically superior to the US rule if covariate information has value. Let $n \equiv \min(N_{xt}, t \in T, x \in X)$ denote the smallest experimental sample. The strong law of large numbers implies that as $n \to \infty$, $W(CS) \to E[\max\{E[y(1)|x], E[y(0)|x]\}]$, which is the optimal mean outcome. Moreover,

$$(36)\quad \lim_{n\to\infty} W(CS) - W(US) \overset{\text{a.s.}}{=} E[\max\{E[y(1)|x], E[y(0)|x]\}] - \max\{E[y(1)], E[y(0)]\}.$$

The right side of (36) is the value of covariate information defined in equation (6). Thus $W(CS) > W(US)$ almost surely if n is sufficiently large and if covariate information has value.

Asymptotic theory may be suggestive, but a planner comparing the CS and US rules must be concerned with their performance in finite samples. An example illustrates the subtlety of the matter:

Example 1: Let the covariate space have two elements, with $X = (a, b)$ and $P(x = a) = P(x = b) = ½$. Let the experimental design be balanced with one subject in each sample, so $N_{a1} = N_{a0} = N_{b1} = N_{b0} = 1$. Let the response distributions $P[y(0)|x = a]$ and $P[y(0)|x = b]$ be degenerate with mass points λ_a and λ_b respectively, where $0 < \lambda_a < 1$, $0 < \lambda_b < 1$, and $1 < \lambda_a + \lambda_b$. Let the response distributions $P[y(1)|x = a]$ and $P[y(1)|x = b]$ be Bernoulli with means μ_a and μ_b respectively, where $0 < \mu_a < 1$ and $0 < \mu_b < 1$.

In this setting, $Q[\bar{y}_{a1} \geq \bar{y}_{a0}] = P[y(1) = 1|x = a] = \mu_a$, $Q[\bar{y}_{b1} \geq \bar{y}_{b0}] = P[y(1) = 1|x = b] = \mu_b$, and $Q[\bar{y}_1 \geq \bar{y}_0] = P[y(1) = 1|x = a]\cdot P[y(1) = 1|x = b] = \mu_a \mu_b$. Hence

$$\begin{aligned}
&W(CS) - W(US)\\
&\quad = ½\{[\mu_a^2 + \lambda_a(1 - \mu_a) + \mu_b^2 + \lambda_b(1 - \mu_b)] - [(\mu_a + \mu_b)\mu_a \mu_b + (\lambda_a + \lambda_b)(1 - \mu_a \mu_b)]\}\\
&\quad = ½\{\mu_a(\mu_a - \lambda_a) + \mu_b(\mu_b - \lambda_b) - \mu_a \mu_b(\mu_a - \lambda_a) - \mu_a \mu_b(\mu_b - \lambda_b)\}
\end{aligned}$$

$$= \tfrac{1}{2}\,\{\mu_a(\mu_a - \lambda_a)(1 - \mu_b) + \mu_b(\mu_b - \lambda_b)(1 - \mu_a)\}.$$

Thus the CS rule is superior or inferior to the US rule, depending on the values of $(\mu_a, \mu_b, \lambda_a, \lambda_b)$. Cases in which $(\mu_a > \lambda_a, \mu_b > \lambda_b)$ are particularly interesting. Then it is optimal is to assign all persons the same treatment, namely $t = 1$. Yet the CS rule, which may assign different treatments to persons with covariates a and b, is superior to the US rule, which enforces uniform treatment assignment. ■

4.2.2. A Sufficient Condition for Superiority of the CS Rule

In general, the treatment-selection probabilities $Q[\bar{y}_{x1} \geq \bar{y}_{x0}]$ and $Q[\bar{y}_1 \geq \bar{y}_0]$ appearing in the expected welfare expressions (34) and (35) are complex functions of the treatment-response distributions $\{P[y(\cdot)|x], x \in X\}$, the sample sizes $(N_{xt}, t \in T, x \in X)$, and the covariate distribution $P(x)$. Hence there seems to be no succinct way to characterize all the circumstances in which one rule outperforms the other. Fortunately, a large-deviations theorem of Hoeffding (1963) for averages of bounded random variables yields relatively simple nonparametric bounds on $Q[\bar{y}_{x1} \geq \bar{y}_{x0}]$ and $Q[\bar{y}_1 \geq \bar{y}_0]$. The result is an instructive sufficient condition for superiority of the CS Rule.

Here is the Hoeffding theorem:

Large Deviations Theorem (Hoeffding, 1963, Theorem 2): Let $w_1, w_2, \cdots, w_n$ be independent real random variables, with bounds $a_i \leq w_i \leq b_i$, $(i = 1, 2, \cdots, n)$. Let $\bar{w} \equiv (1/n)\,\Sigma^n_{i=1} w_i$ and $\mu \equiv E(\bar{w})$. Then, for $v > 0$, $\Pr(\bar{w} - \mu \geq v) \leq \exp[-2n^2v^2/\Sigma^n_{i=1} (b_i - a_i)^2]$. ■

This is a very broad, powerful result. The only distributional assumptions are that the random variables $w_1, w_2, \cdots, w_n$ be independent and have bounded supports. The derived upper bound on $\Pr(\bar{w} - \mu \geq v)$ has no nuisance parameters and is of order $\exp(-nv^2)$ in the sample size n and the distance v. I would note that

Hoeffding (1963), Theorem 1gives tighter but more complicated bounds on $\Pr(\bar{w} - \mu \geq v)$ that hold if $w_1, w_2, \cdots, w_n$ have the same range. It may be that these alternative bounds can be used to improve on the proposition I develop here. I leave this as an open question.

I now apply Hoeffding's Theorem 2 to obtain finite-sample bounds on expected welfare under the CS and US rules. The bounds immediately imply a sufficient condition for superiority of the CS rule. The proposition below requires that the outcome y be bounded but otherwise is entirely general. (I assume that outcomes take values in the unit interval but, given boundedness, this may always be achieved by appropriate normalization of location and scale.) The proof is in an Appendix.

Proposition: Let $T = \{0, 1\}$ and $0 \leq y(t) \leq 1, t \in T$. For $x \in X$, let $M_x \equiv \max\{E[y(1)|x], E[y(0)|x]\}$ and $\delta_x \equiv |E[y(1)|x] - E[y(0)|x]|$. Let $M \equiv \max\{E[y(1)], E[y(0)]\}$ and $\delta \equiv |E[y(1)] - E[y(0)]|$. Then

$$(37) \quad \sum_{x \in X} P(x)\, M_x - \sum_{x \in X} P(x)\delta_x \cdot \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})] \leq W(CS) \leq \sum_{x \in X} P(x)\, M_x.$$

$$(38) \quad M - \delta \cdot \exp[-2\delta^2/\{\Sigma_{x \in X} P(x)^2(N_{x1}^{-1} + N_{x0}^{-1})\}] \leq W(US) \leq M.$$

Hence W(CS) > W(US) whenever

$$(39) \quad \sum_{x \in X} P(x)\delta_x \cdot \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})] < \sum_{x \in X} P(x)\, M_x - M. \quad \blacksquare$$

The right side of (39) is the value of covariate information, which is necessarily non-negative and is positive if optimal treatments vary with x (Section 2.1). The left side of (39) places an upper bound on the divergence of rule CS from the ideal, as defined in (33). This quantity falls to zero as the sample sizes

($N_{x1,}$ N_{x0}; x ∈ X) grow. Hence the proposition reiterates the earlier finding (Section 4.2.1) that the CS rule is superior to the US rule if the samples are sufficiently large and if optimal treatments vary with x. The important new contribution is that its sufficient condition for superiority of the CS rule is a simple explicit function of the sample sizes (N_{x1}, N_{x0}; x ∈ X), the covariate distribution P(x), and the mean treatment responses $\{E[y(\cdot)|x], x \in X\}$. Moreover, this sufficient condition supposes only that outcomes are bounded. No other distributional assumptions are imposed.

Example 2: A numerical example gives a quantitative sense of the proposition. Let X = (a, b), with P(x = a) = P(x = b) = ½. Let the design be balanced, with $N_{a1} = N_{a0} = N_{b1} = N_{b0} = n$, where n is a specified positive integer. Let E[y(1)] = E[y(0)] = ½. Then the CS and US bounds are

$$\tfrac{1}{2}(M_a + M_b) - \tfrac{1}{2}\delta_a \cdot \exp(-n\delta_a^2) - \tfrac{1}{2}\delta_b \cdot \exp(-n\delta_b^2) \leq W(CS) \leq \tfrac{1}{2}(M_a + M_b)$$

$$\tfrac{1}{2} \leq W(US) \leq \tfrac{1}{2}.$$

The table below evaluates the CS bound when E[y(1)|x = a], E[y(0)|x = a], and n have specified values, namely E[y(1)|x = a] = .4, E[y(0)|x = a] ∈ (.4, .5, .6, .7, .8), and n ∈ (1, 10, 25, 50). The quantities E[y(t)|x = b] cannot be varied freely because E[y(t)] = E[y(t)|x = a]P(x = a) + E[y(t)|x = b]P(x = b). Hence the terms of the example require that ½ = ½E[y(t)|x = a] + ½ E[y(t)|x = b], implying that E[y(t)|x = b] = 1 - E[y(t)|x = a]. In the table, treatment 0 is always optimal for persons with x = a and treatment 1 is always optimal for persons with x = b. So covariate information does have value here.

The entries in the column titled "n′" give the smallest value of n such that the lower CS bound exceeds ½, the expected welfare under the US rule; that is, n′ is the smallest integer n such that $(M_a + M_b) - \delta_a \cdot \exp(-n\delta_a^2) - \delta_b \cdot \exp(-n\delta_b^2) > 1$. When n exceeds n′, the CS rule definitely yields higher expected welfare than does the US rule. When n is smaller than n′, the proposition does not yield a ranking of the two rules. The column titled "n″" will be explained in Section 4.3.

The CS Bound

E[y(0)│x = a]	n = 1	n = 10	n = 25	n = 50	n′	n″
.4	[.50, .50]	[.50, .50]	[.50, .50]	[.50, .50]	∞	∞
.5	[.45, .55]	[.46, .55]	[.47, .55]	[.49, .55]	70	196
.6	[.41, .60]	[.47, .60]	[.53, .60]	[.57, .60]	18	48
.7	[.38, .65]	[.53, .65]	[.62, .65]	[.65, .65]	8	20
.8	[.37, .70]	[.62, .70]	[.69, .70]	[.70, .70]	5	5

The first row of the table describes the boundary case in which treatments 0 and 1 yield the same conditional mean outcomes, so the planner is indifferent between the CS and US rules. The other rows show the tension between use of covariate data and sample size. The value of covariate information increases as E[y(0)│x = a] moves away from E[y(1)│x = a], with treatment 0 becoming increasingly better for persons with x = a and, symmetrically, treatment 1 becoming increasingly better for persons with x = b. Hence the upper CS bound increases monotonically. The behavior of the lower CS bound is more complex. As the value of covariate information increases, so does the loss to the planner if covariate data are used to make sub-optimal treatment choices. The result is that, holding sample size fixed, the lower CS bound first falls as E[y(0)│x = a] moves away from E[y(1)│x = a]. In all cases except n = 1, the lower CS bound then rises as E[y(0)│x = a] moves further away.

Although the lower CS bound varies non-monotonically with E[y(0)│x = a], the sample size n′ at which the lower bound first exceeds ½ falls monotonically. Observe how small the values of n′ are. If E[y(0)│x = a] = .5, the CS rule is superior to the US rule in samples of 70 observations or more. If

$E[y(0)|x = a] = .8$, the CS rule is superior in samples of 5 observations or more. ■

4.2.3. Choosing to Use Covariate Information in Practice

A planner deciding whether to use the CS or US rule must face up to the fact that the expected-welfare ranking of the two rules depends on the population distribution of treatment response, which is not known. Computation of the bounds on expected welfare developed in the above proposition requires only knowledge of mean treatment response. The practical problem remains, however, that the planner must make treatment choices using sample data on treatment response, not knowledge of the response distribution.

The inductive nature of inference from samples to population makes it doubtful that there exists any fully satisfactory way to compare STRs such as the CS and US rules in practice. Considering the analogous issue as it arises in point estimation under square loss, Samaniego and Reneau (1994) suppose that the study population is drawn at random from a specified collection of populations, and they evaluate procedures with respect to this superpopulation. However this proposal does not solve the practical problem; now one needs knowledge about the collection of populations from which the study population is drawn.

In the absence of a demonstrably better alternative, I suggest application of standard frequentist thinking to compare the CS and US rules in practice. That is, the planner would use the available sample data to estimate expected welfare under the two rules and then behave as if the estimates are correct. Natural estimates of $\{E[y(1), y(0)|x], x \in X\}$ and $E[y(1), y(0)]$ are $\{(\bar{y}_{x1}, \bar{y}_{x0}), x \in X\}$ and $(\bar{y}_1, \bar{y}_0)$. This suffices to estimate the bounds on expected welfare in (37) and (38). To estimate W(CS) and W(US) themselves requires estimates of the treatment selection probabilities Q[·]. One approach is to use the empirical distribution of the sample data to obtain bootstrap estimates of Q[·], say q[·]. Then the planner would estimate W(CS) and W(US) by

$$(40) \quad w(CS) \equiv \sum_{x \in X} P(x) \{\bar{y}_{x1} \cdot q[\bar{y}_{x1} - \bar{y}_{x0} \geq 0] + \bar{y}_{x0} \cdot q[\bar{y}_{x1} - \bar{y}_{x0} < 0]\}$$

$$(41) \quad w(US) \quad \equiv \; \bar{y}_1 \cdot q[\bar{y}_1 - \bar{y}_0 \geq 0] \; + \; \bar{y}_0 \cdot q[\bar{y}_1 - \bar{y}0 \; < \; 0].$$

If the outcomes y are binary, computation of these bootstrap estimates requires only knowledge of the basic study findings $\{(\bar{y}_{x1}, \bar{y}_{x0}), x \in X\}$ and sample sizes $(N_{x1}, N_{x0}, x \in X)$. However the bootstrap may not be practical in other settings, because studies of treatment response rarely report the full empirical distribution of the observed outcomes. An alternative approach is to base estimates of Q[·] on the classical asymptotic normal approximation to the distribution of $\{(\bar{y}_{x1}, \bar{y}_{x0}), x \in X\}$, which requires only that studies of treatment response report $\{(\bar{y}_{x1}, \bar{y}_{x0}), x \in X\}$ and the standard errors of these estimates.

However one estimates Q[·], frequentist thinking yields a sample-dependent method to choose between the CS and US rules. As such, the method is itself a new STR: a hybrid CS-US rule in which the use of covariate information depends on sample size and on the outcomes observed in the study sample.

4.2.4. Statistical Rules for Treatment of Hypertension

The Materson *et al*. (1993) article on the DVA study of antihypertensive drugs aims to learn how treatment response varies with the race and age of the patient; the authors categorize patients less than 60 years old as "younger" and all others as "older." Table 4 presents the revised findings reported in Figure 2 of Materson, Reda, and Cushman (1995), correcting a computational error in the original article. There are no missing covariate data. The authors perform an intention-to-treat analysis that interprets attrition from the trial as lack of success; from this perspective there are no missing outcome data either. Analyzing the data in Table 4 as the outcomes of a classical randomized experiment, Materson, Reda, and Cushman (1995, p. 189) summarize the findings this way: "Whites responded well to all drug classes, except for lower efficacy of hydrochlorthiazide in younger whites. Blacks responded better to diltiazem than other agents." They do not make specific recommendations for treatment choice but conclude (page 192): "there were important age-by-race differences in success rates that can assist the clinician in selection of an effective drug for an individual patient."

Table 4: Success Rates and Sample Sizes, by Age and Race

Demographic	Treatment													
Group	1		2		3		4		5		6		7	
	$\bar{y}_{x1}$	N_{x1}	$\bar{y}_{x2}$	N_{x2}	$\bar{y}_{x3}$	N_{x3}	$\bar{y}_{x4}$	N_{x4}	$\bar{y}_{x5}$	N_{x5}	$\bar{y}_{x6}$	N_{x6}	$\bar{y}_{x7}$	N_{x7}
a. Younger Blacks	.48	48	.51	35	.43	44	.48	40	.70	37	.42	43	.23	44
b. Older Blacks	.64	44	.45	47	.33	48	.58	45	.85	53	.49	49	.27	44
c. Younger Whites	.32	34	.65	37	.62	39	.69	32	.58	40	.55	33	.26	31
d. Older Whites	.68	60	.72	58	.62	55	.73	60	.72	53	.69	58	.38	64

The data in Table 4 may be used to illustrate application of the expected welfare criterion to choose between the CS and US rules. I shall restrict attention to the two treatments that have the highest success rates for blacks and whites of all ages; treatment 5 (diltiazem) for blacks and treatment 4 (clonidine) for whites. Thus, consider a physician who accepts the DVA success criterion, who observes the race and age of his patients, and who wants to choose treatments in a population where mean treatment response is the same as in the DVA study population. In this setting, the CS rule prescribes that blacks of all ages be assigned treatment 5 and that whites of all ages be assigned treatment 4. The US rule prescription depends on the covariate distribution in the population to be treated. For simplicity, suppose that each of the four demographic groups composes 25 percent of this population. Then $\bar{y}_4 = .62$ and $\bar{y}_5 = .71$, so the US rule prescribes that all patients receive treatment 5.

Using the asymptotic normal approximation to the distribution of $\{(\bar{y}_{x4}, \bar{y}_{x5}), x \in X\}$ yields these estimates of the treatment selection probabilities Q[·]:

$$q[\bar{y}_{a5} - \bar{y}_{a4} \geq 0] = .981; \quad q[\bar{y}_{b5} - \bar{y}_{b4} \geq 0] = .999; \quad q[\bar{y}_{c5} - \bar{y}_{c4} \geq 0] = .166; \quad q[\bar{y}_{d5} - \bar{y}_{d4} \geq 0] = .452$$

$q[\bar{y}_5 - \bar{y}_4 \geq 0] \ = \ .964.$

The resulting estimates of expected welfare are w(CS) = .736 and w(US) = .707. Thus, in this illustration, a physician applying the hybrid CS-US rule of Section 4.2.3 would choose the CS rule.

4.3. Statistical Criteria in Study Design and Interpretation

Empirical research on treatment response has been strongly influenced by the classical theory of hypothesis testing, especially by the idea of testing the null hypothesis of zero average treatment effect in the study population. This hypothesis is institutionalized in the Food and Drug Administration drug approval process, which calls for comparison of the treatment under study (t = 1) with a placebo or an approved treatment (t = 0). Approval of treatment 1 normally requires rejection of the null hypothesis of zero average treatment effect $\{H_0: E[y(1)] = E[y(0)]\}$ in two independent clinical trials (see Fisher and Moyé, 1999). The null hypothesis of zero treatment effect is prominent in experimental design, as researchers use norms for statistical power to choose sample sizes. Moreover, when studies are performed, findings may go unreported or may be deemed to be "insignificant" if they do not meet test-based criteria for statistical precision.

Classical tests of zero average treatment effect play no role in the present analysis of treatment choice using sample data. A planner making treatment choices is concerned with the quantitative variation of outcomes with treatments and covariates. Testing the null hypothesis of zero effect does not address this question. Indeed, the absence of a well-defined decision problem motivating classical testing theory has long formed a central part of the Bayesian critique of frequentist statistics.

It would be of much interest to reconsider the present FDA drug approval process and current norms for experimental design from the treatment-choice perspective of this paper. These are serious matters and I shall refrain from making hasty recommendations here. I would, however, reiterate the recommendation on collection and reporting of covariate information made in Section 2.2. That is, whenever there is reason to think that treatment response may be heterogeneous, researchers designing studies should aim to collect extensive covariate information on subjects and to report findings on mean treatment response conditional on these covariates.

It is easy to demonstrate the negative consequences for treatment choice of conventional statistical criteria for reporting covariate information. Consider the experimental setting of Section 4.2. Findings on mean treatment response conditional on covariates often go unreported if the null hypothesis $\{H_0: E[y(1)|x] = E[y(0)|x]\}$ is not rejected. In particular, researchers often use the t-statistic criterion

$$(42) \quad \text{Report } (\bar{y}_{x1}, \bar{y}_{x0}) \text{ if } (\bar{y}_{x1} - \bar{y}_{x0})/[\text{SVar }(\bar{y}_{x1} - \bar{y}_{x0})]^{½} > 2,$$

where SVar $(\bar{y}_{x1} - \bar{y}_{x0})$ is the conventional sample estimate of the variance of $(\bar{y}_{x1} - \bar{y}_{x0})$. If findings conditional on covariates are not reported, planners are denied the opportunity to apply the CS rule. In fact, planners may not be able to apply even the US rule. Computation of the unconditional success rates $\bar{y}_t$, $t \in T$ requires knowledge of the success rates conditional on covariates if the distribution of covariates in the population to be treated differs from that in the study population.

Example 3: The example in Section 4.2.2 gives a quantitative sense of the implications of conventional reporting criteria. Consider the idealized t-statistic criterion

$$(43) \quad \text{Report } (\bar{y}_{x1}, \bar{y}_{x0}) \text{ if } E_Q(\bar{y}_{x1} - \bar{y}_{x0})/[\text{Var}_Q(\bar{y}_{x1} - \bar{y}_{x0})]^{½} > 2.$$

I refer to this as an "idealized" criterion because the operational t-statistic rule given in (42) makes reporting a function of the sample drawn, hence a random variable, whereas the idealized t-statistic given in (43) makes reporting a function of population characteristics specified in the example. Let y be binary, so $E[y(\cdot)|x] = P[y(\cdot) = 1|x]$. Let $P_{tx} \equiv P[y(t) = 1|x]$, $t = 0, 1$. Then the idealized criterion becomes

(44) Report $(\bar{y}_{x1}, \bar{y}_{x0})$ if $(P_{1x} - P_{0x})/[P_{1x}(1 - P_{1x})/n + P_{0x}(1 - P_{0x})/n]^{1/2} > 2$.

The column titled "n″" in the table of Section 4.2.2 gives the minimal value of n at which this reporting criterion is met. Comparison of the entries for n′ and n″ shows that $n' \le n''$ in every case and that n″ is much larger than n′ when $P_{0x} \in (.5, .6, .7)$. Thus, use of a reporting criterion based on statistical precision may prevent use of the CS rule when that rule is superior to the US rule. ■

5. Conclusion

The objective of informing treatment choice provides an explicit practical motivation for empirical study of treatment response. This paper has shown that empirical research seeking to inform treatment choices differs in important respects from analyses that aim to discover "causal effects" or to perform classical hypothesis tests. Adopting the treatment-choice perspective systematically affects how one should select a study population, cope with identification problems, and make use of finite-sample data.

We have examined a class of treatment-choice problems – a social planner choosing treatments to maximize mean outcome in a heterogeneous population – that is familiar and analytically tractable. There clearly are other treatment-choice problems that warrant investigation. For example, the planner may face budgetary or other constraints under which optimal treatments turn out to differ from those that maximize mean outcomes conditional on covariates. Or the planner may want to maximize something other than the mean population outcome, perhaps the median. There is considerable scope for future work investigating how studies of treatment response may inform treatment choice in a variety of settings of practical interest.

Whatever specific features the treatment-choice problem may have, problems of identification and statistical inference must be confronted. I would reiterate that these are logically distinct issues, identification being a matter of deduction of properties of populations and statistical inference being one of induction from samples to populations. A fundamental open question is how a decision maker choosing treatments should integrate the distinct forms of ambiguity generated by identification problems and statistical inference.

Appendix: Proof of the Proposition

CS Bound: The upper bound follows from (34), so the task is to prove the lower bound. For x ∈ X, I write $\bar{y}_{x1} - \bar{y}_{x0}$ as the average of independent random variables and apply Hoeffding's theorem to show that

$$\text{(A1)}\quad M_x - \delta_x\cdot\exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})] \le E[y(1)|x]\cdot Q[\bar{y}_{x1} - \bar{y}_{x0} \ge 0] + E[y(0)|x]\cdot Q[\bar{y}_{x1} - \bar{y}_{x0} < 0].$$

Let $N_x \equiv N_{x1} + N_{x0}$. For each x ∈ X,

$$\text{(A2)}\quad \bar{y}_{x1} - \bar{y}_{x0} = (1/N_{x1}) \sum_{j \in N(x,1)} y_j - (1/N_{x0}) \sum_{j \in N(x,0)} y_j$$
$$= (1/N_x)[\sum_{j \in N(x,1)} (y_j\cdot N_x/N_{x1}) + \sum_{j \in N(x,0)} (-y_j\cdot N_x/N_{x0})].$$

Thus $\bar{y}_{x1} - \bar{y}_{x0}$, the difference between the averages of N_{x1} and N_{x0} random variables with range [0, 1], can be reformulated as the average of N_x random variables; the first N_{x1}have range $[0, N_x/N_{x1}]$ and the remaining N_{x0} have range $[-N_x/N_{x0}, 0]$.

Consider x ∈ X such that E[y(1)|x] < E[y(0)|x]. Then $E(\bar{y}_{x1} - \bar{y}_{x0}) = -\delta_x$. Hoeffding's theorem yields

$$\text{(A3)}\quad Q[\bar{y}_{x1} - \bar{y}_{x0} \ge 0] \le \exp[-2N_x^2\,\delta_x^2/\{N_{x1}\cdot(N_x/N_{x1})^2 + N_{x0}\cdot(N_x/N_{x0})^2\}] = \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})].$$

Hence

$$\text{(A4)}\quad E[y(1)|x]\cdot Q[\bar{y}_{x1} - \bar{y}_{x0} \ge 0] + E[y(0)|x]\cdot Q[\bar{y}_{x1} - \bar{y}_{x0} < 0]$$
$$\ge E[y(1)|x]\cdot\exp[-2\delta_x^2 / (N_{x1}^{-1} + N_{x0}^{-1})] + E[y(0)|x]\cdot\{1 - \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})]$$
$$= M_x - \delta_x\cdot\exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})].$$

So (A1) holds. Next consider x ∈ X such that E[y(1)|x] > E[y(0)|x]. For such x, $E(\bar{y}_{x0} - \bar{y}_{x1}) = -\delta_x$. Application of Hoeffding's theorem yields

(A5) $Q[\bar{y}_{x1} - \bar{y}_{x0} < 0] = Q[\bar{y}_{x0} - \bar{y}_{x1} > 0] \le Q[\bar{y}_{x0} - \bar{y}_{x1} \ge 0] \le \exp[-2\delta_x^2/(N_{x1}^{-1} + N_{x0}^{-1})]$.

Thus (A1) continues to hold by an argument analogous to (A4). Finally consider $x \in X$ such that $E[y(1)|x] = E[y(0)|x]$. For such x, $\delta_x = 0$. Hence (A1) holds as an equality.

<u>US Bound</u>: The upper bound follows from (35). The lower bound holds as an equality if E[y(1)] = E[y(0)]. The task is to show that the lower bound holds otherwise. As in the proof of the CS bound, I write $\bar{y}_1 - \bar{y}_0$ as the average of independent random variables and then apply Hoeffding's theorem.

Let $N \equiv \Sigma_{x \in X} (N_{x1} + N_{x0})$. Then

(A6) $\bar{y}_1 - \bar{y}_0 = \sum_{x \in X} P(x) \ (1/N_{x1}) \sum_{j \in N(x, 1)} y_j - \sum_{x \in X} P(x) \ (1/N_{x0}) \sum_{j \in N(x, 0)} y_j$

$= (1/N)[\sum_{x \in X} \sum_{j \in N(x, 1)} (y_j \cdot P(x)N/N_{x1}) + \sum_{x \in X} \sum_{j \in N(x, 0)} (-y_j \cdot P(x)N/N_{x0})]$.

Thus $\bar{y}_1 - \bar{y}_0$ averages N independent random variables with ranges $[0, P(x)N/N_{x1}]$ and $[-P(x)N/N_{x0}, 0]$, $x \in X$.

Let E[y(1)] < E[y(0)]. Then $E(\bar{y}_1 - \bar{y}_0) = -\delta$. Application of Hoeffding's theorem yields

(A7) $Q[\bar{y}_1 - \bar{y}_0 \ge 0] \le \exp[-2N^2 \delta^2/\{\Sigma_{x \in X} N_{x1}(P(x) \cdot N/N_{x1})^2 + N_{x0}(P(x) \cdot N/N_{x0})^2\}]$

$= \exp[-2\delta^2/\{\Sigma_{x \in X} P(x)^2(N_{x1}^{-1} + N_{x0}^{-1})\}]$.

Hence

(A8) $E[y(1)] \cdot Q[\bar{y}_1 - \bar{y}_0 \ge 0] + E[y(0)] \cdot Q[\bar{y}_1 - \bar{y}_0 < 0] \ge M - \delta \cdot \exp[-2\delta^2/\{\Sigma_{x \in X} P(x)^2(N_{x1}^{-1} + N_{x0}^{-1})\}]$.

The same result holds when E[y(1)] > E[y(0)].

Q. E. D.

References

Angrist, J., G. Imbens, and D. Rubin (1996), “Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91, 444-455

Balke, A. and J. Pearl(1997), "Bounds on Treatment Effects from Studies With Imperfect Compliance," Journal of the American Statistical Association, 92, 1171-1177.

Berger, J. (1985), Statistical Decision Theory and Bayesian Analysis, New York: Springer-Verlag.

Campbell, D. (1984), “Can We Be Scientific in Applied Social Science?” Evaluation Studies Review Annual, 9, 26-48.

Campbell, D. and R. Stanley (1963), Experimental and Quasi-Experimental Designs for Research, Chicago: Rand McNally.

Crits-Christoph, P. Siqueland, L., Blaine, J., Frank, A., et. al. (1999), “Psychosocial treatments for cocaine dependence,” Archives of General Psychiatry, 56, 493-502.

Dehejia, R. (1999), “Program Evaluation as a Decision Problem,” National Bureau of Economic Research Working Paper 6954.

Doolittle, F. and J. Riccio (1992), “Case Management in Welfare Employment Programs,” in C. Manski and I. Garfinkel (eds.), Evaluating Welfare and Training Programs, Cambridge, MA: Harvard University Press.

Fisher, L. and L. Moyé (1999), “Carvedilol and the Food and Drug Administration Approval Process: An Introduction,” Controlled Clinical Trials, 20, 1-15.

Freis, E.D., Materson, B.J., and Flamenbaum, W. (1983), “Comparison of Propranolol or Hydorchlorothiazide Alone for Treatment of Hypertension, III: Evaluation of the Renin-Angiotensin System,” The American Journal of Medicine, 74, 1029-1041.

Heckman, J., J. Smith, and N. Clements (1997), "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts,” Review of Economic Studies, 64, 487-535.

Heckman, J. and E. Vytlacil (2000), “Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect,” Discussion Paper T0259, National Bureau of Economic Research.

Hoeffding, W. (1963), “Probability Inequalities for Sums of Bounded Random Variables,” Journal of the American Statistical Association, 58, 13-30.

Horowitz, J. and C. Manski (1998), “Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations,” Journal of Econometrics, 84, 37-58.

Horowitz, J., and C. Manski (2000), “Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data,” Journal of the American Statistical Association, 95, 77-84.

Hotz, J., C. Mullins, and S. Sanders (1997), "Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analyzing the Effects of Teenage Childbearing," Review of Economic Studies, 64, 575-603.

Imbens, G. and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62, 467-476.

Little, R. (1992), "Regression with Missing X's: A Review," Journal of the American Statistical Association, 87, 1227-1237.

Manski, C. (1989), "Anatomy of the Selection Problem," Journal of Human Resources, 24, 343-360.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," American Economic Review Papers and Proceedings, 80, 319-323.

Manski, C. (1995), Identification Problems in the Social Sciences, Cambridge: Harvard University Press.

Manski, C. (1997), "Monotone Treatment Response," Econometrica, 65, 1311-1334.

Manski, C. (2000), "Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice," Journal of Econometrics, 95, 415-442.

Manski, C. and D. Nagin (1998), "Bounding Disagreements About Treatment Effects: A Case Study of Sentencing and Recidivism," Sociological Methodology, 28, 99-137.

Manski, C. and J. Pepper (2000), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," Econometrica, 68, 997-1010.

Materson, B.J., Reda, D.J., Cushman, W.C., Massie, B.M., Freis, E.D., Kochar, M.S., Hamburger, R.J., Fye, C., Lakshman, R., Gottdiener, J., Ramirez, E.A., and Henderson, W.G. (1993), "Single-Drug Therapy for Hypertension in Men: A Comparison of Six Antihypertensive Agents with Placebo," The New England Journal of Medicine, 328, 914-921.

Materson, B.J., Reda, D.J. and Cushman, W.C. (1995). "Department of Veterans Affairs Single-Drug Therapy of Hypertension Study: Revised Figures and New Data," American Journal of Hypertension, 8, 189-192.

Pepper, J. (2000), "What Do Welfare-to-Work Demonstrations Reveal to Welfare Reformers?" Thomas Jefferson Center Discussion Paper, University of Virginia.

Robins, J. (1989), "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in Sechrest, L., H. Freeman, and A. Mulley. eds. Health Service Research Methodology: A Focus on AIDS, NCHSR, U.S. Public Health Service.

Robins, J. and S. Greenland (1996), "Comment on Angrist, Imbens, and Rubin's 'Identification of Causal Effects Using Instrumental Variables'," Journal of the American Statistical Association, 91, 456-458.

Robins, J., Rotnitzky, A., and Zhao, L. (1994), "Estimation of Regression Coefficients When Some

Regressors Are Not Always Observed," Journal of the American Statistical Association, 89, 846-866.

Rosenbaum, P. (1999), "Choice as an Alternative to Control in Observational Studies," Statistical Science, 14, 259-304.

Samaniego, F. and D. Reneau (1994), "Toward a Reconciliation of the Bayesian and Frequentist Approaches to Point Estimation," Journal of the American Statistical Association, 89, 947-957.

Wald, A. (1950), Statistical Decision Functions, New York: Wiley.

Wang, C., Wang, S., Zhao, L. and Ou, S. (1997), "Weighted Semiparametric Estimation in Regression Analysis with Missing Covariate Data," Journal of the American Statistical Association, 92, 512-525.