# Treatment Choice with Trial Data:
# Statistical Decision Theory Should Supplant Hypothesis Testing

Charles F. Manski

Department of Economics and Institute for Policy Research
Northwestern University, Evanston, IL 60208 USA

Revised: August 2018
forthcoming in *The American Statistician*

Abstract

A central objective of empirical research on treatment response is to inform treatment choice. Unfortunately, researchers commonly use concepts of statistical inference whose foundations are distant from the problem of treatment choice. It has been particularly common to use hypothesis tests to compare treatments. Wald's development of statistical decision theory provides a coherent frequentist framework for use of sample data on treatment response to make treatment decisions. A body of recent research applies statistical decision theory to characterize uniformly satisfactory treatment choices, in the sense of maximum loss relative to optimal decisions (also known as maximum regret). This paper describes the basic ideas and findings, which provide an appealing practical alternative to use of hypothesis tests. For simplicity, the paper focuses on medical treatment with evidence from classical randomized clinical trials. The ideas apply generally, encompassing use of observational data and treatment choice in non-medical contexts.

keywords: analysis of treatment response, randomized clinical trials, minimax regret, medical decisions

author footnote: Charles F. Manski is Board of Trustees Professor, Department of Economics, Northwestern University, Evanston, IL 60208 USA. cfmanski@northwestern.edu

acknowledgments: I am grateful to the Editors and reviewers for their comments.

# 1. INTRODUCTION

A central objective of empirical research on treatment response is to inform treatment choice. Identification problems combine with the necessity of inference from sample data to limit the informativeness of studies. Unfortunately, researchers commonly use concepts of statistical inference whose foundations are distant from the problem of treatment choice. It has been particularly common to use hypothesis tests to compare treatments.

The Wald (1950) development of statistical decision theory provides a coherent frequentist framework for use of sample data on treatment response to make treatment decisions. A body of recent research applies statistical decision theory to characterize uniformly satisfactory treatment choices, in the sense of maximum loss relative to optimal decisions (also known as maximum regret). This paper describes the basic ideas and findings, which provide an appealing practical alternative to use of hypothesis tests.

To keep the exposition simple and framed in a familiar setting, I focus on medical treatment with evidence from classical randomized clinical trials. Trials have long enjoyed a favored status within evidence-based medicine, often being called the "gold standard" for collection of data on treatment response. The influential Cochrane system for grading the quality of evidence ordinarily reserves its highest rating for evidence from randomized trials (Higgins and Green, 2011, Sec. 12.2.1). The drug approval process of the U.S. Food and Drug Administration (FDA) ordinarily considers only experimental evidence when making decisions on drug approval. While I focus on the use of trial data in medical decision making, the broad ideas discussed here are general, encompassing use of observational data and treatment choice in non-medical contexts.

Section 2 reviews the use of hypothesis tests to compare medical treatments and the basic principles of statistical decision theory. Section 3 describes the recent research on uniformly satisfactory treatment choice using trial data. Section 4 concludes.

Readers may be aware that it has become increasingly common to express concern that evaluation of empirical research by the outcome of hypothesis tests generates publication bias and diminishes the reproducibility of findings. See, for example, Ioannidis (2005) and Wasserstein and Lazar (2016). This concern is important, but it is distinct from the theme of the present paper.

The paper relates directly to other important issues covered in the *ASA Statement on Statistical Significance and P-Values* (Wasserstein and Lazar, 2016). Two the six Principles of the Statement are:

> "3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold."
>
> "5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result."

Treatment choice using statistical decision theory is not based at all on whether a p-value passes a threshold. Statistical decision theory clearly distinguishes between the statistical and clinical significance of empirical estimates of treatment effects.

## 2. BACKGROUND

### 2.1. Using Hypothesis Tests to Compare Treatments

A longstanding practice in medicine has been to use trial data to test a specified null hypothesis against an alternative and to use the outcome of the test to compare treatments. A common procedure when comparing two treatments in a trial is to view one as the status quo and the other as an innovation. The usual null hypothesis is that the innovation is no better than the status quo and the alternative is that the innovation is better. If the null hypothesis is not rejected, it is recommended that the status quo

treatment continue to be used in clinical practice. If the null is rejected, it is recommended that the innovation become the treatment of choice. This type of test is institutionalized in the FDA drug approval process, which calls for comparison of a new drug with a placebo or a previously approved treatment. Approval of the new drug normally requires rejection of the null hypothesis of zero average treatment effect in two independent trials (Fisher and Moyé, 1999).

The convention has been to perform a test that fixes the probability of rejecting the null hypothesis when it is correct, the probability of a Type I error. Then sample size determines the probability of rejecting the alternative hypothesis when it is correct, the probability of a Type II error. The power of a test is defined as one minus the probability of a Type II error. The convention has been to choose a sample size that yields specified power at some value of the effect size deemed clinically important. For example, International Conference on Harmonisation (1999) has provided guidance for the design and conduct of trials evaluating pharmaceuticals, stating (p. 1923):

> "Conventionally the probability of type I error is set at 5% or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results. The probability of type II error is conventionally set at 10% to 20%."

Manski and Tetenov (2016) observe that there are several reasons why hypothesis testing may yield unsatisfactory results for medical decisions and other forms of treatment choice. These include:

*1. Use of Conventional Asymmetric Error Probabilities*: It has been standard to fix the probability of Type I error at 5% and the probability of Type II error at 10-20%. The theory of hypothesis testing gives no rationale for selection of these conventional error probabilities. It gives no reason why a clinician concerned with patient welfare should find it reasonable to make treatment choices that have a substantially greater probability of Type II than Type I error.

*2. Inattention to Magnitudes of Losses to Welfare When Errors Occur*: A clinician should care about more than the probabilities of Type I and II error. He should care as well about the magnitudes of the losses to patient welfare that arise when errors occur. A given error probability should be less acceptable when the welfare difference between treatments is larger, but the theory of hypothesis testing does not take this into account.

*3. Limitation to Settings with Two Treatments*: A clinician often chooses among several treatments and many clinical trials compare more than two treatments. Yet the standard theory of hypothesis testing only contemplates choice between two treatments. Statisticians have struggled to extend it to deal sensibly with comparisons of multiple treatments.

The third issue is well-appreciated, but the first two are often overlooked. A simple example shows why they may matter for patient care.

2.1.1. Example

Suppose that a typically terminal form of cancer may be treated by a status quo treatment or an innovation. It is known from experience that mean patient life span with the status quo treatment is one year. Prior to use of the innovation, medical researchers see two possibilities for its effectiveness. It may be less effective than the status quo, yielding a mean life span of only 1/3 of a year, or it may be much more effective, yielding a mean life span of 5 years.

Suppose that a classical randomized trial is performed to learn the effectiveness of the innovation. Let the trial data be used to perform a conventional hypothesis test comparing the innovation and the status quo. The null hypothesis is that the innovation is no more effective than the status quo and the alternative is that the innovation is more effective. The probability of a Type I error is set at 0.05 and that of a Type II error is 0.20. The test result is used to choose between the treatments.

A Type I error occurs with frequentist probability 0.05 and reduces mean patient life span by 2/3 of a year (1 year minus 1/3 year). A Type II error occurs with frequentist probability 0.20 and reduces mean patient life span by 4 years (5 years minus 1 year). Thus, use of the test to choose between the status quo and the innovation implies that society is willing to tolerate a large (0.20) chance of a large welfare loss (4 years) when making a Type II error, but only a small (0.05) chance of a small welfare loss (2/3 of a year) when making a Type I error. The theory of hypothesis testing does not motivate this asymmetry.

## 2.2. Principles of Statistical Decision Theory

### 2.2.1. Basic Ideas

The standard formalization of decision making under uncertainty supposes that a decision maker must choose among a set of feasible actions. The welfare achieved by an action depends on an unknown feature of the environment, called the *state of nature*. The decision maker wants to choose an action that maximizes welfare, but he can't do this with certainty because the state of nature is unknown. The decision maker lists all states of nature that he believes could possibly occur. This list, the *state space*, expresses partial knowledge. The larger the state space, the less the decision maker knows about the outcomes of actions.

For example, the decision maker may be a clinician and the actions may be treatments for a patient. A state of nature may characterize how a patient would respond to alternative treatments, which may be incompletely known. Welfare may be a health outcome of interest, perhaps patient life span or quality of life, that would occur when a specified treatment is administered to a patient. The clinician might ideally want to choose a treatment that optimizes the patient's health outcome but, having incomplete knowledge of treatment response, he can't do this with certainty.

Suppose that a sampling process generates observable sample data that are informative about the

true state; for example, data on how each member of a sample of patients has responded to the assigned treatment. Wald (1950) considered the general problem of using such sample data to make decisions. He posed the task as choice of a *statistical decision function*, which maps potentially available data into a choice among the feasible actions. Wald's seminal book is abstract, making it a difficult read. Ferguson (1967), Berger (1985), and Parmigiani and Inoue (2009) provide comprehensive expositions.

Wald recommended ex ante evaluation of statistical decision functions as *procedures* applied as the sampling process is engaged repeatedly to draw independent data samples. The idea of a procedure transforms the original statistical problem of induction from a single sample into the deductive problem of assessing the probabilistic performance of a statistical decision function across realizations of the sampling process. Thus, the theory is frequentist.

Wald proposed that the decision maker evaluate a statistical decision function by the mean welfare it yields across realizations of the sampling process. His presentation differed semantically from the one that I use to describe treatment choice in that he defined loss to be the negative of welfare, took the objective to be minimization of loss rather than maximization of welfare, and used the term *risk* to denote mean loss across realizations of the sampling process. With these semantic distinctions, he prescribed a three-step decision process:

(1) Specify the set of feasible actions, the loss (or welfare) function, and the state space. These basic concepts of decision theory are context specific. The set of feasible actions is commonly considered to be predetermined. The loss function and the state space are subjective. The former formalizes what the decision maker wants to achieve and the latter expresses what states of nature he believes could possibly occur.

(2) Eliminate inadmissible statistical decision functions. A decision function is inadmissible if there exists another that yields at least as good mean sampling performance in every state of nature and strictly better mean performance in some state.

(3) Use some criterion to choose an admissible statistical decision function. Wald focused on the

*minimax* criterion and on minimization of a subjective mean of the risk function (called *Bayes risk*). Savage (1951) proposed *minimax regret*.

2.2.2. Decision Criteria

What are reasonable ways to choose an admissible decision function? The Bayesian approach is particularly well known. It has been common to think of the Bayesian process of transforming a prior into a posterior distribution as antithetical to frequentist statistics, but Wald provided a clear frequentist perspective on Bayes decisions. He showed that minimization of Bayes risk, a frequentist decision criterion, yields the same decisions as would occur if one performs Bayesian inference, combining the prior distribution with the data to form a posterior subjective distribution, and then chooses an action to minimize the posterior mean of expected loss. Berger (1985), Section 4.4.1 gives an accessible proof.

Bayesian decision making is compelling when one feels able to place a credible subjective prior distribution on the state space. There exists a considerable body of work ranging across multiple disciplines that develops methods to help persons conceptualize uncertainty and express themselves in subjective probabilistic terms. See, for example, Savage (1971), Koriat, Lichtenstein, and Fischhoff (1980), Morgan and Henrion (1990), Manski (2004), and Garthwaite, Kadane, and O'Hagan (2005).

Nevertheless, Bayesians have long struggled to provide guidance on specification of priors to be used in evidence-based medicine and the matter continues to be controversial. See, for example, the spectrum of views regarding Bayesian analysis of randomized trials expressed by the authors and discussants of Spiegelhalter, Freedman, and Parmar (1994). The controversy suggests that inability to express a credible prior is common in actual decision settings.

When one finds it difficult to assert a credible subjective distribution, Bayesian statisticians who believe it essential to use a probability distribution to express uncertainty may suggest use of some default distribution that is variously called a “reference” or “conventional” or “objective” prior; see, for example, Berger (2006). However, there is no consensus on what prior should play this role. The choice made

matters for decision making.

An alternative is to abandon the notion that one must use a probability distribution to express uncertainty. In the absence of a prior, Wald argued that a reasonable way to act is to use a decision criterion that achieves uniformly satisfactory results, whatever the true state of nature may be. There are multiple ways to formalize the idea of uniformly satisfactory results. The two most commonly studied are embodied in the minimax and minimax-regret (MR) criteria.

The minimax criterion chooses an action that minimizes the maximum risk that might possibly occur, across all feasible states of nature. The minimax-regret criterion considers each state of nature and computes the incremental risk that occurs if one chooses a specified action rather than the one that minimizes risk in this state. This quantity, called *regret*, measures the nearness to optimality of the specified action in the state of nature. The decision maker must choose without knowing the true state. To achieve a uniformly satisfactory result, he computes the maximum regret of each action; that is, the maximum distance from optimality that the action would yield across all possible states of nature. The MR criterion chooses an action that minimizes maximum regret.

The minimax and MR criteria are sometimes confused with one another, but they yield the same choice only in certain special cases. Whereas the minimax criterion considers only the worst outcome that an action may yield, MR considers the worst outcome relative to what is achievable in a given state of nature. Savage (1951) distinguished the minimax criterion sharply from MR, writing that the former criterion is "ultrapessimistic" while the latter is not. Maximum regret quantifies how uncertainty--lack of knowledge of the true state of nature--potentially diminishes the quality of decisions.

It is important to understand that use of the minimax or the MR criteria does not eliminate all subjectivity in decision making. As discussed earlier, decision theory begins with specification of a welfare function and a state space, both of which are subjective. Bayes decision theory goes a step further by placing a subjective distribution on the state space. The minimax and MR criteria do not embrace this further element of subjectivity, but they still require the decision maker to specify a welfare

function and state space.

The reader may have noticed that, to introduce this discussion, I asked what are "reasonable" decision criteria rather than what is an "optimal" criterion. Statistical decision theorists recognized from the outset that there is no singularly optimal way to choose an admissible decision function. There at most are reasonable ways. Wald (1950), who was particularly concerned with decision making in the absence of a prior distribution on the state space, motivated his focus on the minimax criterion in part by stating (p. 18):

> "a minimax solution seems, in general, to be a reasonable solution of the decision problem when an a priori distribution . . . . does not exist or is unknown to the experimenter."

Ferguson (1967) wrote (p. 28):

> "It is a natural reaction to search for a 'best' decision rule, a rule that has the smallest risk no matter what the true state of nature. Unfortunately, *situations in which a best decision rule exists are rare and uninteresting*. For each fixed state of nature there may be a best action for the statistician to take. However, this best action will differ, in general, for different states of nature, so that no one action can be presumed best overall."

He went on to write (p. 29): "A *reasonable* rule is one that is better than just guessing."

2.2.3. Some History, Post Wald

The Wald framework has breathtaking generality. In principle, it enables comparison of all statistical decision functions whose risk functions exist. It enables comparison of alternative sampling processes as well as decision rules. It uses no asymptotic approximations. It applies whatever information the decision maker may have. The state space may be finite dimensional or larger; that is, nonparametric. The true state of nature may be point or partially identified. Settings with partial identification are ones where the sampling process generating the data incompletely reveals the true state asymptotically; see Manski (2003, 2007a).

Given the appeal of statistical decision theory, one might anticipate that it would play a central role in modern statistics. However, this has not occurred. After publication of Wald (1950), a surge of important extensions and applications followed in the 1950s. Much research focused on best point prediction under square loss with sample data, analysis of which began with Hodges and Lehmann (1950). In this important case, regret is mean square error and the MR criterion yields a predictor that minimizes maximum mean square error across the state space.

However, this period of rapid development closed by the 1960s, with the exception of Bayesian statistical decision theory. Bayesian analysis has continued to develop, but as a self-contained field of study disconnected from the Wald frequentist framework. Recent research in Bayesian statistics has focused more on the computational problem of transformation of priors into posteriors than on use of posteriors in decision making.

Why did statistical decision theory lose momentum long ago? One reason may have been the technical difficulty of the subject. Wald's ideas are easy to describe abstractly, but applying them can be analytically and computationally demanding. Determination of admissible decision functions and minimax/minimax-regret rules is often difficult. Another reason may have been diminishing interest in decision making as the motivation for analysis of sample data. Modern statisticians tend to view their objectives as estimation and hypothesis testing rather than decision making.

I cannot be sure what role these or other reasons played in the vanishing of statistical decision theory from statistics in the latter part of the twentieth century. However, the near absence of the subject in mainstream journals and textbooks of the period is indisputable. I think this is unfortunate. The recent research described in the next sections aims to reinvigorate statistical decision theory, focusing on the important applied problem of treatment choice. Other recent research, not described here, provides new analysis of best point prediction under square loss when the sampling process is afflicted with missing data or other problems of imperfect data quality; see Dominitz and Manski (2017).

# 3. RECENT WORK ON STATISTICAL DECISION THEORY FOR TREATMENT CHOICE

Bayesian statistical decision theory has long been available to design trials and to choose treatments with trial data. DeGroot (1970) provides a classical treatise on the subject. Canner (1970), Spiegelhalter, Freedman, and Parmar (1994), Cheng, Su, and Berry (2003), and Spiegelhalter (2004) study various aspects. However, as mentioned above, Bayesians have struggled to provide guidance on specification of priors and the matter continues to be controversial. Perhaps as a result, Bayesian analysis is well known but seldom used in evidence-based medicine. A limited exception is that the FDA has provided guidance permitting the use of Bayesian statistics in the design and analysis of clinical trials evaluating new medical devices; see U. S. Food and Drug Administration (2010).

I describe here a recent body of research that avoids specification of priors and instead studies uniformly satisfactory treatment choice, using maximum regret to measure performance across states of nature. Contributions to this emerging literature include Manski (2004, 2005, 2007a, 2007b), Schlag (2006), Hirano and Porter (2009), Stoye (2009, 2012), Tetenov (2012), Manski and Tetenov (2016), and Kitagawa and Tetenov (2018).

The recent research supposes that the objective of treatment choice is to maximize a social welfare function that sums treatment outcomes across a population of patients that may have heterogeneous treatment response. For example, the objective may be to maximize the five-year survival rate in a population of cancer patients or mean life span in a population with a chronic disease. In this setting, a statistical decision function uses the data to choose an allocation of patients to treatments. Using terminology introduced in Manski (2004), such a function has been called a *statistical treatment rule* (*STR*).

The mean sampling performance of an STR across repeated samples is its *expected welfare.* This term means the negative of Wald's risk and is used because the objective is to maximize welfare rather than minimize loss. Given that the objective is to maximize rather than minimize a function, the

minimax decision criterion becomes maximin instead. The MR criterion remains as earlier except that regret in a state of nature now is the maximum welfare achievable in that state minus the expected welfare of a specified STR.

In what follows, Section 3.2 explains why the recent work has measured the performance of STRs by maximum regret. Section 3.3. discusses treatment choice with existing trial data. Section 3.4 considers the design of trials.

## 3.2. Measuring Performance by Maximum Regret

In the absence of a prior distribution on the state space, practical and conceptual reasons motivate measurement of the performance of STRs by maximum regret across the state space, rather than by minimum expected welfare. I explain here.

### 3.2.1. Practical Appeal

From a practical perspective, it has been found that MR decisions behave more reasonably than do maximin ones in the context of treatment choice. In common settings of treatment choice with trial data on outcomes that take a bounded range of values, it has been found that the MR rule is well approximated by the *empirical success* (ES) rule, which chooses the treatment with the highest observed average outcome in the trial. The ES rule provides a simple and plausible way to use the results of a trial. The performance of the ES rule from the perspective of maximum regret was initiated by Manski (2004). Subsequently, Schlag (2006) and Stoye (2009) showed that this rule either exactly or approximately minimizes maximum regret in common settings with two treatments when sample size is moderate. Hirano and Porter (2009) showed that the ES rule is asymptotically optimal.

In contrast, the maximin rule commonly ignores the trial data, whatever they may be. When Savage (1951) stated that the minimax criterion is "ultrapessimistic," he went on to write (p. 63): "it can

lead to the absurd conclusion in some cases that no amount of relevant experimentation should deter the actor from behaving as though he were in complete ignorance." Savage did not flesh out this statement but it is easy to show that this occurs with trial data. Manski (2004) provides a simple example that will be discussed in Section 3.3.

### 3.2.2. Conceptual Appeal

The conceptual appeal of using maximum regret to measure performance is that maximum regret quantifies how lack of knowledge of the true state of nature diminishes the quality of decisions. While the term "maximum regret" has become standard in the literature, it is important to keep in mind that this term is a shorthand for the maximum sub-optimality of a decision criterion across the feasible states of nature. An STR with small maximum regret is uniformly near-optimal across all states.

Maximum regret is well-defined in general settings with multiple treatments and when patients have heterogeneous observable covariates that may be used to differentiate treatment. However, the concept is especially transparent when there are two treatments and the members of the patient population are observationally identical, all having the same observable covariates.

Suppose there are two feasible treatments, say A and B. In a state of nature where A is better, the regret of an STR is the product of the probability across repeated samples that the rule commits a Type I error (choosing B) and the magnitude of the loss in expected welfare that occurs when choosing B. Similarly, in a state where B is better, regret is the probability of a Type II error (choosing A) times the magnitude of the loss in expected welfare when choosing A.

Recall the critique in Section 2.1 of the conventional use of hypothesis testing to choose a treatment. I called attention to the asymmetric treatment of Type I and Type II error probabilities and the inattention to magnitudes of losses when errors occur. Evaluating treatment rules by regret overcomes both problems. Regret considers Type I and II error probabilities symmetrically and it measures the magnitudes of the losses that errors produce.

3.2.3. Example

To illustrate, consider again the example of Section 2.1.1, in which a conventional hypothesis test is used as an STR to choose between a status quo treatment for cancer and an innovation. There are two feasible states of nature in the example, with the innovation yielding mean life span of 1/3 year in one state and 5 years in the other. In the first state, the regret of this conventional "test rule" equals 1/30 of a year; that is, a 0.05 chance of a Type I error times a 2/3 of a year reduction in mean life span with improper choice of the innovation. In the second state, the regret of the test rule equals 4/5 of a year; that is, a 0.20 chance of a Type II error times a 4-year reduction in mean life span with improper choice of the status quo. Thus, the maximum regret of the test rule is 4/5 of a year.

Rather than use the conventional test rule to choose between the status quo and the innovation, one could seek an STR that has smaller maximum regret. Given the available trial data, a simple option would be to reverse the conventional probabilities of Type I and Type II; thus, one might use a test with a 0.20 chance of a Type I error and a 0.05 chance of a Type II error. In the first state, the regret of this unconventional test rule STR equals 2/15 of a year; that is, a 0.20 chance of a Type I error times a 2/3 of a year reduction in mean life span with improper choice of the innovation. In the second state, the regret of the unconventional test rule equals 1/5 of a year; that is, a 0.05 chance of a Type II error times a 4-year reduction in mean life span with improper choice of the status quo. Thus, the maximum regret of the unconventional test rule is 1/5 of a year.

In this example, the unconventional test rule delivers much smaller maximum regret than does the conventional test rule. There may exist other STRs that perform even better.

3.3. Treatment Choice with Existing Trial Data

To move beyond verbal discussion and examples, I now formalize treatment choice as a statistical

decision problem in the relatively simple setting of a classical trial with two treatments and a population of observationally identical patients. The presentation in this section draws substantially on Manski (2007, Chapter 12). The research literature cited earlier also studies more general and complex settings with multiple treatments, patients who have heterogeneous observable covariates, and imperfect trials that only partially identify treatment response.

3.3.1. General Analysis

Suppose that a health planner must assign treatment A or B to each member of patient population J. Each patient $j \in J$ has response function $y_j(\cdot): T \to Y$ mapping treatments $t \in T$ into individual outcomes $y_j(t) \in R$. Let P denote the distribution of treatment response in the population.

The members of the population may respond heterogeneously to treatment, but they are observationally identical to the planner. For any $\delta \in [0, 1]$, the planner can allocate a fraction $\delta$ of patients to treatment B and $1 - \delta$ to A. The planner wants to choose $\delta$ to maximize an additive welfare function

$$U(\delta, P) = E[y(A)]\cdot(1 - \delta) + E[y(B)]\cdot\delta \;=\; \alpha\cdot(1 - \delta) + \beta\cdot\delta \;=\; \alpha + (\beta - \alpha)\cdot\delta, \qquad (1)$$

where $\alpha \equiv E[y(A)]$ and $\beta \equiv E[y(B)]$ are the mean outcomes if everyone were to receive treatment A or B respectively. The quantity $\beta - \alpha$ is the average treatment effect (ATE) in the population. It is optimal to set $\delta = 1$ if the ATE is positive and $\delta = 0$ if the ATE is negative. The problem of interest is treatment choice when incomplete knowledge of P makes it impossible to determine the sign of the ATE.

Suppose that sample data are available. Let Q be the sampling distribution and $\Psi$ be the sample space. For example, the data may be treatment response observed in a randomized trial. A statistical treatment rule (STR) is a function $\delta(\cdot): \Psi \to [0, 1]$ that maps sample data into a treatment allocation. The welfare realized with $\delta$ and data $\psi$ is the random variable

$$U(\delta, P, \psi) = \alpha + (\beta - \alpha)\cdot\delta(\psi). \quad (2)$$

The state space $[(P_s, Q_s), s \in S]$ is the set of (P, Q) pairs that the planner deems possible. Expected welfare in state s, the mean sampling performance of rule $\delta$ in this state, is

$$W(\delta, P_s, Q_s) = \alpha_s + (\beta_s - \alpha_s)\cdot E_s[\delta(\psi)]. \quad (3)$$

Here $E_s[\delta(\psi)] \equiv \int_\Psi \delta(\psi)dQ_s(\psi)$ is the expected allocation of patients to treatment B, across repeated samples.

Rule $\delta$ is admissible if there exists no rule $\delta'$ such that $W(\delta', P_s, Q_s) \geq W(\delta, P_s, Q_s)$ for all $s \in S$ and $W(\delta', P_s, Q_s) > W(\delta, P_s, Q_s)$ for some s. The Bayes, maximin, and MR rules are as follows:

*Bayes rule*: $\max_{\delta \in [0, 1]} \int_S W(\delta, P_s, Q_s)d\pi(s),$ (4)

where $\pi$ is a subjective distribution on the state space.

*Maximin rule*: $\max_{\delta \in [0, 1]} \min_{s \in S} W(\delta, P_s, Q_s).$ (5)

*Minimax-regret rule*: $\min_{\delta \in [0, 1]} \max_{s \in S} [\max(\alpha_s, \beta_s) - W(\delta, P_s, Q_s)].$ (6)

3.3.2. Illustration: Choice Between a Status Quo Treatment and an Innovation when Outcomes are Binary

To illustrate in perhaps the simplest non-trivial setting, let the outcomes y be binary, taking the value zero if treatment fails and one if it succeeds. Let A be a status quo treatment and B be an

innovation. Suppose that the planner knows the success probability $\alpha \equiv P[y(A) = 1]$ of the status quo treatment but not the success probability $\beta \equiv P[y(B) = 1]$ of the innovation. The planner wants to choose treatments to maximize the success probability.

A randomized trial is performed to learn about outcomes under the innovation, with N subjects randomly drawn from the population and assigned to treatment B. The observed trial outcomes are that n subjects realize outcome y = 1 and N − n realize y = 0. In this setting, N indexes the sampling process and the number n of experimental successes is a sufficient statistic for the data.

The feasible STRs are functions $\delta(\cdot)$: [0, . . . , N] ⇾ [0, 1] that map the number of experimental successes into a treatment allocation. The expected welfare of rule δ is

$$W(\delta, P, N) = \alpha + (\beta - \alpha)\cdot E[\delta(n)]. \tag{7}$$

n is distributed binomial **B**[β, N], so

$$E[\delta(n)] = \sum_{i=0}^{N} \delta(i)\cdot f(n = i; \beta, N), \tag{8}$$

where $f(n = i; \beta, N) \equiv N![i!\cdot(N - i)!]^{-1}\beta^i(1 - \beta)^{N-i}$ is the probability of i successes. The only unknown determinant of expected welfare is β, so the state space S indexes the feasible values of β. Specifically, $\beta_s \equiv P_s[y(b) = 1]$.

It is reasonable in this setting to conjecture that admissible treatment rules should be ones in which the fraction of the population allocated to treatment B increases with n. It turns out that the admissible treatment rules are a simple subclass of these rules. A theorem of Karlin and Rubin (1956) shows that the admissible rules are the *monotone treatment rules*. Monotone rules assign all persons to the status quo if the experimental success rate is below some threshold and all to the innovation if the

success rate is above the threshold. Thus, $\delta$ is admissible if and only if

$$\delta(n) = 0 \quad \text{for } n < n_0, \tag{9a}$$

$$\delta(n) = \lambda \quad \text{for } n = n_0, \tag{9b}$$

$$\delta(n) = 1 \quad \text{for } n > n_0, \tag{9c}$$

for some $0 \leq n_0 \leq N$ and $0 \leq \lambda \leq 1$.

The collection of monotone treatment rules is a mathematically "small" subset of the space of all feasible treatment rules. Nevertheless, it still contains a broad range of rules. These include

*Data-Invariant Rules*: These are the rules $\delta(\cdot) = 0$ and $\delta(\cdot) = 1$, which assign all persons to treatment a or b respectively, whatever n may be.

*Empirical Success Rule*: An optimal treatment rule allocates all persons to treatment A if $\beta < \alpha$ and all to B if $\beta > \alpha$. The empirical success rule emulates the optimal rule by replacing $\beta$ with its sample analog, the empirical success rate n/N.

*Bayes Rules*: The form of the Bayes rule depends on the prior subjective distribution placed on $\beta$. Consider the class of Beta priors, which form the conjugate family for a Binomial likelihood. Let $(\beta_s, s \in S) = (0, 1)$ and let the prior be Beta with parameters (c, d). Then the posterior mean for $\beta$ is $(c + n)/(c + d + N)$. The resulting Bayes rule is

$$\delta(n) = 0 \quad \text{for } (c + n)/(c + d + N) < \alpha, \tag{10a}$$

$$\delta(n) = \lambda \quad \text{for } (c + n)/(c + d + N) = \alpha, \text{ where } 0 \leq \lambda \leq 1, \tag{10b}$$

$$\delta(n) = 1 \quad \text{for } (c + n)/(c + d + N) > \alpha. \tag{10c}$$

*Maximin Rule*: Minimum expected welfare for rule δ is

$$\min_{s \in S} W(\delta, P_s, N) = \alpha + \min_{s \in S} (\beta_s - \alpha) E_s[\delta(n)]. \quad (11)$$

$E_s[\delta(n)] > 0$ for all $\beta_s > 0$ and for all monotone rules except $\delta(\cdot) = 0$. When S contains states with $\beta_s < \alpha$, the maximin rule is $\delta(\cdot) = 0$. Thus, the maximin rule ignores the trial data on treatment response, whatever they may turn out to be. This illustrates the Savage (1951) statement that using the maximin rule can induce one to entirely ignore available sample data.

*Minimax-Regret Rule*: The regret of rule δ in state s is

$$\max(\alpha, \beta_s) - \{\alpha + (\beta_s - \alpha) \cdot E_s[\delta(n)]\} \quad (12)$$
$$= (\beta_s - \alpha)\{1 - E_s[\delta(n)]\} \cdot 1[\beta_s \geq \alpha] + (\alpha - \beta_s) E_s[\delta(n)] \cdot 1[\alpha \geq \beta_s].$$

Thus, regret is the mean welfare loss when a member of the population is assigned the inferior treatment, multiplied by the expected fraction of the population assigned this treatment. The minimax-regret rule does not have an analytical solution, but it can be determined numerically. When all values of β are feasible, the minimax-regret rule is well approximated by the empirical success rule.

3.3.3. Numerical Computation of the Minimax-Regret Rule for Small Samples

Manski (2007, Table 12.1) reports numerical computations of the minimax-regret rule for specified values of α and N when all values of β are feasible; that is, when $(\beta_s, s \in S) = [0, 1]$. I reproduce the findings here in Table 1. The top two panels display the value of $(n_0, \lambda)$ for this rule. The third panel displays the value of minimax regret.

The top panel of the table shows that the threshold $n_0$ of experimental successes for allocation of persons to treatment B increases with the sample size and with the success probability of treatment A. The inequality $|n_0 - \alpha N| \leq 1$ holds everywhere in the table. Thus, the minimax-regret rule is well approximated by an empirical success rule.

The third panel shows that the value of minimax regret decreases by roughly an order of magnitude as the sample size increases from 0 to 10. For example, when $\alpha = 0.50$, it falls from 0.25 to 0.027. Thus, even a sample size as small as 10 suffices to make maximum regret quite small.

TABLE 1: MINIMAX-REGRET RULES FOR SMALL SAMPLE SIZES

| $n_0$: threshold sample size | N = 0 | N = 1 | N = 2 | N =3 | N = 4 | N = 5 | N = 6 | N = 7 | N = 8 | N = 9 | N = 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| α = 0.10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| α = 0.25 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| α = 0.50 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 |
| α = 0.75 | 0 | 1 | 2 | 2 | 3 | 4 | 5 | 5 | 6 | 7 | 8 |
| α = 0.90 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 8 | 9 |

| $\lambda$: threshold allocation | N = 0 | N = 1 | N = 2 | N = 3 | N = 4 | N = 5 | N = 6 | N = 7 | N = 8 | N = 9 | N = 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| α = 0.10 | 0.9 | 0.67 | 0.52 | 0.41 | 0.32 | 0.26 | 0.18 | 0.09 | 0 | 0.89 | 0.78 |
| α = 0.25 | 0.75 | 0.36 | 0.17 | 0.93 | 0.67 | 0.42 | 0.18 | 0.93 | 0.67 | 0.43 | 0.18 |
| α = 0.50 | 0.5 | 1 | 0.5 | 1 | 0.5 | 1 | 0.5 | 1 | 0.5 | 1 | 0.5 |
| α = 0.75 | 0.25 | 0.64 | 0.83 | 0.07 | 0.33 | 0.58 | 0.82 | 0.07 | 0.33 | 0.57 | 0.82 |
| α = 0.90 | 0.1 | 0.33 | 0.48 | 0.59 | 0.68 | 0.74 | 0.82 | 0.91 | 1 | 0.11 | 0.22 |

| minimax regret value | N = 0 | N = 1 | N = 2 | N = 3 | N = 4 | N = 5 | N = 6 | N = 7 | N = 8 | N = 9 | N = 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| α = 0.10 | 0.09 | 0.067 | 0.052 | 0.041 | 0.033 | 0.027 | 0.022 | 0.019 | 0.017 | 0.017 | 0.017 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| α = 0.25 | 0.19 | 0.09 | 0.052 | 0.039 | 0.038 | 0.035 | 0.03 | 0.027 | 0.027 | 0.025 | 0.023 |
| α = 0.50 | 0.25 | 0.063 | 0.063 | 0.044 | 0.044 | 0.035 | 0.035 | 0.03 | 0.03 | 0.027 | 0.027 |
| α = 0.75 | 0.19 | 0.09 | 0.052 | 0.039 | 0.038 | 0.035 | 0.03 | 0.027 | 0.027 | 0.025 | 0.023 |
| α = 0.90 | 0.09 | 0.067 | 0.052 | 0.041 | 0.033 | 0.027 | 0.022 | 0.019 | 0.017 | 0.017 | 0.016 |

Source: Manski (2007), Table 12.1.

### 3.4. Designing Trials to Enable Near-Optimal Treatment Choice

From the perspective of treatment choice, an ideal objective for the design of trials would be to collect data that enable subsequent implementation of an optimal treatment rule in the patient population of interest; that is, a rule for use of trial data that always selects the best treatment, with no chance of error. Optimality is too strong a property to be achievable with finite sample size. However, near-optimal rules---ones with small maximum regret---exist when classical trials are large enough.

Manski and Tetenov (2016) investigate trial design that enables near-optimal treatment choices. It is shown that, given any $\varepsilon > 0$, *ε-optimal* rules exist when trials have large enough sample size. An $\varepsilon$-optimal rule has expected welfare, across repeated samples, within $\varepsilon$ of the welfare of the best treatment in every state of nature. Equivalently, it has maximum regret no larger than $\varepsilon$.

The article considers trials that draw predetermined numbers of subjects at random within groups stratified by covariates and treatments. It reports exact results for cases of two treatments and binary outcomes; see Section 3.4.1 below for a summary. It gives sufficient conditions on sample sizes that ensure existence of $\varepsilon$-optimal treatment rules when there are multiple treatments and outcomes are bounded. These conditions are obtained by application of large deviations inequalities to evaluate the performance of empirical success rules.

Choosing sample size to enable existence of $\varepsilon$-optimal treatment rules requires specification of a value for ε. The selected ε determines how much deviation from optimality a decision maker is willing to tolerate when making treatment choices. The value of ε should be specified by clinical researchers concerned with patient care rather than by some universal convention. Clinical researchers may, perhaps, find it congenial to let ε equal the *minimum clinically important difference* (MCID) in the average treatment effect comparing alternative treatments.

Medical research has long distinguished between the statistical and clinical significance of treatment effects. While the idea of clinical significance has been interpreted in various ways, many writers call an average treatment effect clinically significant if its magnitude is greater than a specified value deemed minimally consequential in clinical practice. International Conference on Harmonisation (1999) put it this way (p. 1923): "The treatment difference to be detected may be based on a judgment concerning the minimal effect which has clinical relevance in the management of patients."

### 3.4.1. Findings with Binary Outcomes, Two Treatments, and Balanced Designs

Determination of sample sizes that enable near-optimal treatment is simple in settings with binary outcomes (coded 0 and 1 for simplicity), two treatments, and a balanced design which assigns the same number of subjects to each treatment group. Manski and Tetenov (2016, Table 1), reproduced here as Table 2, provides exact computations of the minimum sample size that enables $\varepsilon$-optimality when a clinician uses one of three different treatment rules, for various values of $\varepsilon$.

The first column shows the minimum sample size (per treatment arm) that yields $\varepsilon$-optimality when a clinician uses the empirical success (ES) rule to make a treatment decision. The ES rule chooses the treatment with the better average outcome in the trial. The rule assigns half the population to each treatment if there is a tie. It is known that the ES rule minimizes maximum regret in settings with binary outcomes, two treatments, and balanced designs (Stoye, 2009).

The second and third columns display the minimum sample sizes that yield $\varepsilon$-optimality of rules

based on one-sided 5% and 1% hypothesis tests. Decisions made with these tests take the two treatments to be a status quo and an innovation, choosing the innovation if the estimated treatment effect is positive and statistically significant. There is no consensus on what hypothesis test should be used to compare two proportions. Results are reported based on the widely used one-sided two-sample z-test, which is based on an asymptotic normal approximation (Fleiss, 1973).

The findings are remarkable. A sample as small as 2 observations per treatment arm makes the ES rule $\varepsilon$-optimal when $\varepsilon = 0.1$ and a sample of size 145 suffices when $\varepsilon = 0.01$. The minimum sample sizes required for $\varepsilon$-optimality of the test rules are orders of magnitude larger. If the z-test of size 0.05 is used, a sample of size 33 is required when $\varepsilon = 0.1$ and 3488 when $\varepsilon = 0.01$. The sample sizes must be more than double these values if the z-test of size 0.01 is used. See Manski and Tetenov (2016) for discussion of the factors that underlie these findings.

Table 2: Minimum Sample Sizes per Treatment Enabling ε-Optimal Treatment Choice: Binary Outcomes, Two Treatments, Balanced Designs

| $\varepsilon$ | ES Rule | One-Sided 5% z-Test | One-Sided 1% z-Test |
|---|---|---|---|
| 0.01 | 145 | 3488 | 7963 |
| 0.03 | 17 | 382 | 879 |
| 0.05 | 6 | 138 | 310 |
| 0.10 | 2 | 33 | 79 |
| 0.15 | 1 | 16 | 35 |

Source: Manski and Tetenov (2016), Table 1.

3.4.2. Implications for Practice

Based on their exact calculations and analytical findings using large-deviations inequalities,

Manski and Tetenov conclude that sample sizes determined by clinically relevant near-optimality criteria tend to be much smaller than ones set by conventional statistical power criteria. Reduction of sample size relative to prevailing norms can be beneficial in multiple ways. Reduction of total sample size can lower the cost of executing trials, the time necessary to recruit adequate numbers of subjects, and the complexity of managing trials across multiple centers. Reduction of sample size per treatment arm can make it feasible to perform trials that increase the number of treatment arms and, hence, yield information about a wider variety of treatment options.

## 4. CONCLUSION

Science does not always progress monotonically. There are times when important, even fundamental, ideas are discovered and receive attention but then are neglected. This has occurred with statistical decision theory, which received considerable attention in the middle of the twentieth century but was largely forgotten by the century's end. A revival in the context of treatment choice began in the early 2000s and has been gathering force. I hope that the growing dissatisfaction of statisticians with ritual applications of hypothesis testing, exemplified by the ASA Statement in Wasserstein and Lazar (2016), will encourage statisticians to relearn statistical decision theory and use it when studying not only treatment choice with trial data but decision making with sample data more generally.

References

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer: New York.

Berger, J. (2006), "The Case for Objective Bayesian Analysis," *Bayesian Analysis*, 1, 385-402.

Canner, P. (1970), "Selecting One of Two Treatments When the Responses Are Dichotomous," *Journal of the American Statistical Association*, 65, 293-306.

Cheng, Y., F. Su, and D. Berry (2003), "Choosing Sample Size for a Clinical Trial Using Decision Analysis," *Biometrika*, 90, 923-936.

DeGroot, M. (1970), *Optimal Statistical Decisions*, New York: McGraw-Hill.

Dominitz, J. and C. Manski (2017), "More Data or Better Data? A Statistical Decision Problem," *Review of Economic Studies*, 84, 1583-1605.

Ferguson, T. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press: San Diego.

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

Fleiss, J. (1973), *Statistical Methods for Rates and Proportions*, New York: Wiley.

Garthwaite, P., J. Kadane, and A. O'Hagan (2005), "Statistical Methods for Eliciting Probability Distributions," *Journal of the American Statistical Association*, 100, 680-701.

Hodges, E. and E. Lehmann (1950), "Some Problems in Minimax Point Estimation," *Annals of Mathematical Statistics*, 21, 182-197.

International Conference on Harmonisation (1999) ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonized tripartite guideline. *Statistics in Medicine,* 18, 1905-1942.

Ioannidis, J. (2005), "Why Most Published Research Findings are False," *PLoS Medicine*, 2, 696-701.

Kitagawa, T. and A. Tetenov (2018), "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591-616.

Koriat, A., S. Lichtenstein, and B. Fischhoff (1980), "Reasons for Confidence," *Journal of Experimental Psychology: Human Learning and Memory*, 6, 107-118.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer.

Manski, C. (2004a), "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 221-246.

Manski, C. (2004), "Measuring Expectations," *Econometrica*, 72, 1329-1376.

Manski, C. (2005), *Social Choice with Partial Knowledge of Treatment Response*, Princeton: Princeton University Press.

Manski, C. (2007a), *Identification for Prediction and Decision,* Cambridge: Harvard University Press.

Manski, C. (2007b), "Minimax-Regret Treatment Choice with Missing Outcome Data," *Journal of Econometrics*, 139, 105-115.

Manski, C. and A. Tetenov (2016), "Sufficient Trial Size to Inform Clinical Practice," Proceedings of the *National Academy of Sciences*, 113, 10518-10523.

Morgan, G. and M. Henrion (1990), *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, New York: Cambridge University Press.

Parmigiani, G. and L. Inoue (2009), *Decision Theory: Principles and Approaches*, New York: Wiley.

Savage, L. (1951), "The Theory of Statistical Decision," *Journal of the American Statistical Assocation,* 46, 55-67.

Savage, L. (1971), "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association*, 66, 783-801.

Schlag, K. (2006), "Eleven–Tests Needed for a Recommendation," European University Institute Working Paper ECO No. 2006/2.

Spiegelhalter D., L. Freedman, and M. Parmar (1994), "Bayesian Approaches to Randomized Trials" (with discussion), *Journal of the Royal Statistics Society Series A*, 157, 357-416.

Spiegelhalter, D. (2004), "Incorporating Bayesian Ideas into Health-Care Evaluation," *Statistical Science*, 19, 156-174.

Stoye, J. (2009), "Minimax Regret Treatment Choice with Finite Samples," *Journal of Econometrics*, 151, 70-81.

Stoye, J. (2012), "Minimax Regret Treatment Choice with Covariates or with Limited Validity of Experiments," *Journal of Econometrics*, 166, 138-156.

U.S. Food and Drug Administration (2010), *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*, https://www.fda.gov/MedicalDevices/ucm071072.htm, accessed July 20, 2018.

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

Wasserstein, R. and N. Lazar (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129-133.