TRIAL SIZE FOR NEAR-OPTIMAL CHOICE BETWEEN SURVEILLANCE AND AGGRESSIVE TREATMENT: Reconsidering MSLT-II

Charles F. Manski
Department of Economics and Institute for Policy Research, Northwestern University

and

Aleksey Tetenov
Department of Economics, University of Bristol

Revised: October 2018
Forthcoming, *The American Statistician*

Abstract

A convention in designing randomized clinical trials has been to choose sample sizes that yield specified statistical power when testing hypotheses about treatment response. Manski and Tetenov (2016) critiqued this convention and proposed enrollment of sufficiently many subjects to enable *near-optimal* treatment choices. This paper develops a refined version of the 2016 analysis that is applicable to trials comparing aggressive treatment of patients with surveillance. The need for a refined analysis arises because the earlier work assumed that there is only a primary health outcome of interest, without secondary outcomes. An important aspect of choice between surveillance and aggressive treatment is that the latter may have side effects. One should then consider how the primary outcome and side effects jointly determine patient welfare. This requires new analysis of sample design. As a case study, we reconsider a trial comparing nodal observation and lymph node dissection when treating patients with cutaneous melanoma. Using a statistical power calculation, the investigators assigned 971 patients to dissection and 968 to observation. We conclude that assigning 244 patients to each option would yield findings that enable suitably near-optimal treatment choice. Thus, a much smaller sample size would have sufficed to inform clinical practice.

We are grateful to the editor and reviewers for constructive comments.

1. Introduction

A core objective of randomized clinical trials is to inform treatment choice. The convention has been to choose sample sizes that yield specified statistical power when testing designated null hypotheses about treatment response against designated alternatives. However, statistical power and hypothesis testing are at most loosely connected to effective patient care.

Manski and Tetenov (2016) critiqued the use of power calculations to set sample size and developed an alternative principle that aims to inform patient care directly. They proposed enrollment of sufficiently many subjects to enable determination of near-optimal treatment choices and provided criteria to accomplish this. An optimal treatment rule would always select the best treatment, with no chance of error. This is infeasible to achieve with trials having finite sample size. *Near-optimal* treatment choices are ones that are suitably close to this ideal. The article gave numerical calculations of sufficient sample sizes for trials with binary outcomes.

A broad conclusion was that sample sizes determined by clinically relevant near-optimality criteria are much smaller than ones set using conventional statistical power calculations. A variety of factors contribute to this conclusion. One is that the near-optimality perspective considers Type I and Type II errors symmetrically. In contrast, power calculations are performed with the probability of a Type I error a priori constrained to equal 0.05 or another conventional value. Another is that the near-optimality perspective evaluates products of effect sizes and error probabilities. It allows larger error probabilities if the two treatments are nearly equivalent from a patient welfare perspective than if one treatment is substantially better for patients.

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 need to perform 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 more treatment options.

This paper develops and applies a refined version of the analysis in Manski and Tetenov (2016) to trials that compare aggressive treatment of patients with surveillance. Patient care abounds with instances of this choice problem. Internists choose between prescription of pharmaceuticals and surveillance when treating patients at risk of heart disease or diabetes. Oncologists choose between surveillance and aggressive treatments such as surgery or chemotherapy when treating cancer patients at risk of metastasis. Aggressive treatment may be appealing to the extent that it better prevents onset or reduces the severity of illness. Surveillance may be attractive to the extent that it avoids side effects that may occur with aggressive treatment.

The need for a refined version of the analysis in Manski and Tetenov (2016) arises because this earlier work studied settings in which there is only a primary health outcome of interest, without any secondary outcomes. An important aspect of choice between surveillance and aggressive treatment is that the latter may have side effects. The prevailing approach to choice of sample size in trials has been to focus entirely on the primary outcome of a treatment, without considering secondary outcomes. This practice is reasonable when the primary outcome is the dominant determinant of patient welfare or, put another way, when there is little variation in secondary outcomes across treatments. It is not reasonable otherwise. When the side effects of aggressive treatments are serious, it is more reasonable to consider how the primary outcome and side effects jointly determine patient welfare. This requires new analysis of sample design, regardless of whether one adopts the perspective of near optimality or statistical power.

To provide a realistic case study illuminating the general issues, we consider a trial comparing *nodal observation* and *lymph node dissection* (*lymphadenectomy*) when treating patients with early-stage

cutaneous melanoma at risk of metastasis. Nodal observation is surveillance of lymph nodes by ultrasound scan, a procedure that has negligible side effects. Lymph node dissection is a surgical procedure in which the lymph nodes in the relevant regional basin are removed. Dissection is commonly viewed as an aggressive treatment. A particularly concerning side effect is chronic swelling in the region of lymph node removal, a condition called lymphedema, which may reduce patient quality of life substantially (Cheville *et al.*, 2010). Choice between nodal observation and lymph node dissection is a common decision faced in early treatment of melanoma, breast cancer, and other forms of localized cancer. We focus on melanoma because there has long been controversy about the merits of dissection relative to observation in this context. See Faries (2018).

The Multicenter Selective Lymphadenectomy Trial II (MSLT-II) compared dissection and observation for melanoma patients who had recently undergone sentinel lymph-node biopsy and who had obtained a positive finding of malignancy. The primary outcome was defined to be melanoma-specific survival for three years following the date of randomization. Findings were reported in Faries *et al*. (2017).

Our concern is choice of sample size in the trial. Using a statistical power calculation, the investigators assigned 971 patients to dissection and 968 to observation. Considering sample size from the perspective of near-optimal treatment choice, we conclude that assigning 244 patients to each treatment would yield findings that enable suitably near-optimal treatment choice. Thus, a much smaller sample size would have sufficed to inform clinical practice.

We perform the computations under the assumption that treatment choice will be made with the *empirical success* rule, which selects the treatment with the higher sample average welfare. When used to choose between two treatments, the empirical success rule approximately yields equal Type I and Type II error probabilities. This contrasts sharply with conventional hypothesis tests, which yield asymmetric error probabilities (typically 0.05 for Type I errors and 0.10 to 0.20 for Type II errors). The empirical success

rule provides a simple and plausible way to use the results of a trial. Analysis of its performance from the perspective of near optimality was initiated by Manski (2004). Stoye (2009) shows that use of this rule either exactly or approximately minimizes the sample size needed to achieve near optimality in common settings with two treatments.

The analyses in Manski and Tetenov (2016) and in the present paper focus on trials such as MSLT-II that are performed to help clinicians choose between treatments that are already used in practice, rather than for a regulatory purpose such as drug approval. Trials performed for regulatory purposes may be analyzed from the perspective of improving patient welfare, but the analysis may need to recognize legal constraints and the incentives of the firms or other entities that perform the trials. See Tetenov (2016) for discussion of these matters.

2. Methods

2.1. Critique of the Use of Power Calculations to Set Trial Size

The use of statistical power calculations to set trial size derives from the presumption that outcome data will be used to test a null hypothesis against an alternative. A common practice is to use a hypothesis test to recommend which of two treatments, say A and B, should be given to a patient population. The null hypothesis often is that treatment B is no better than A, and the alternative is that B is better. If the null hypothesis is not rejected, choice of A is recommended. If the null is rejected, B is recommended.

The standard practice has been to perform a test that fixes at a predetermined level the probability of rejecting the null hypothesis when it is correct (a Type I error). Then sample size determines the probability

of rejecting the alternative hypothesis when it is correct (a Type II error). The power of a test is 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 average treatment effect deemed clinically relevant. For example, International Conference on Harmonisation (1999) provides guidance for the design 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%."

Trials with samples too small to achieve these error probabilities are called "underpowered" and are criticized as scientifically useless and medically unethical (e.g., Halpern, Karlawish, and Berlin, 2002).

MSLT-II followed the standard practice of setting sample size to obtain statistical power. The investigators enrolled 1939 subjects, citing this reasoning (Faries, *et al*., 2017, page 2214): "We estimated that with a total sample of 1925 patients, the trial would have a power of 83% to detect a between-group difference of 5 percentage points in melanoma-specific survival." Specifically, the investigators contemplated using the trial data to perform a two-tailed hypothesis test. The null hypothesis is that both treatments yield the same rate of three-year melanoma-specific survival. The alternative is that the survival rates differ by at least 5 percentage points. The probabilities of Type I and Type II errors are 0.05 and 0.17. The study investigators provide detailed statements of the study protocol and statistical analysis plan in a supplementary online document.[1]

Manski and Tetenov (2016) critique the standard practice, observing that there are several reasons why hypothesis testing may yield unsatisfactory results for medical decisions. These include

---

[1] www.nejm.org/doi/suppl/10.1056/NEJMoa1613210/suppl_file/nejmoa1613210_protocol.pdf

Use of conventional asymmetric error probabilities: It has been standard to set 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 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.

Disregard of magnitudes of losses 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.

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 long struggled to extend it to deal sensibly with comparisons of multiple treatments, without consensus on how to do this.

See Manski (2018) for further critique of the use of hypothesis tests to make treatment choices.

2.2. Setting Sample Size to Enable Near-Optimal Treatment

An ideal objective for trial design would be to collect data that enable optimal treatment choice in the patient population of interest, with no chance of error. Optimality is too strong a property to be achievable

with finite sample size, but near-optimal rules exist when trials with perfect internal and external validity are large enough.

Near optimality was suggested as a principle for decision making under uncertainty by Savage (1951) within an essay commenting on the seminal Wald (1950) development of statistical decision theory. Statistical decision theory studies the broad problem of decision making when one has incomplete knowledge of the "state of nature" (that is, how patients respond to treatment), but can collect informative sample data. For example, a clinician who has incomplete knowledge of treatment response may use trial data to become better informed.

Considering each possible state of nature, Savage proposed computation of the mean loss in welfare that would occur across repeated samples if one were to choose a specified treatment rule rather than the one that is best in this state. This quantity measures the nearness to optimality of the specified treatment rule in each state of nature. The actual decision problem requires choice of a treatment rule without knowing the true state of nature. The decision maker can evaluate a rule by the maximum distance from optimality that it yields across all possible states of nature. He can then choose a rule that minimizes the maximum distance from optimality. Doing so yields a rule that is uniformly near optimal. Statistical decision theory has used the term *regret* as a shorthand for nearness to optimality. Maximum nearness to optimality is called *maximum regret* and the rule that minimizes the maximum distance from optimality is called the *minimax-regret* rule. See Manski (2018) for further discussion of the Wald theory and minimax regret.

The concept of near optimality is applicable in general settings with multiple treatments, but it is easiest to explain when there are two treatments, say A and B. In states of nature where A is better, the nearness to optimality of a specified treatment rule 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 patient welfare that occurs when choosing B. Similarly, in states where B is better, nearness to optimality is the probability

of a Type II error (choosing A) times the magnitude of the loss in welfare when choosing A. In contrast to the use of hypothesis testing to choose a treatment, evaluation by near optimality views Type I and II error probabilities symmetrically and it assesses the magnitudes of the welfare losses that errors produce.

Manski and Tetenov (2016) investigated trial design that enables near-optimal treatment. They supposed that the objective is to maximize average welfare across the relevant patient population. For example, the objective may be to maximize the five-year survival rate of a population of cancer patients or the average number of quality-adjusted life years of patients with a chronic disease.

They considered trials that draw predetermined numbers of subjects at random within groups stratified by treatments and observed patient risk factors. They showed that, given any specified positive value of a constant ε measuring nearness to optimality, *ε-optimal* rules exist when trials have large enough sample size. A ε-optimal rule is one whose mean value of average patient welfare, across repeated samples, is within ε of the optimum in every state of nature.

They reported exact numerical results for the case of two treatments with binary outcomes such as survival versus death. They gave simple sufficient conditions on sample sizes that ensure existence of ε-optimal treatment rules when there are multiple treatments and outcomes are bounded.

2.3. Choosing the MSLT-II Sample Size to Enable Near-Optimal Treatment

The sample-size calculations in Manski and Tetenov (2016) concerned settings where treatments do not have side effects. Hence, they should not be applied to MSLT-II or other trials comparing surveillance and aggressive treatment. The present paper develops a refined version of the earlier analysis that recognizes the possibility of side effects.

We assume a simple patient welfare function that transparently expresses patient concern with both survival, the primary outcome of treatment, and lymphedema, a possible secondary outcome. Patients may perhaps have more complex welfare functions than the one posed here. Our methodology for setting sample size may be applied with any welfare function. The specific findings depend on the welfare function.

Let welfare with nodal observation equal 1 if a patient survives for three years and equal 0 otherwise. Welfare with dissection depends on whether a patient experiences lymphedema. When a patient does not experience lymphedema, welfare with dissection equals 1 if the patient survives for three years and equals 0 otherwise. When a patient experiences lymphedema, welfare is lowered by a specified fraction h, whose value expresses the harm associated with lymphedema. Thus, a patient who experiences lymphedema has welfare $1 - h$ if he survives and $-h$ if he does not survive.

When making the treatment decision, a clinician does not know whether a given patient will survive and/or experience lymphedema. To cope with uncertainty about patient-specific treatment response, medical economists have recommended that clinicians maximize average welfare in a relevant patient population. Implementation of this recommendation does not require that a clinician knows patient-specific treatment response. It does, however, require that the clinician know average treatment response in the patient population. See, for example, Phelps and Mushlin (1988) and Meltzer (2001).

A few symbols help to explain in the setting of choice between nodal observation and dissection. Let nodal observation be treatment A. Let y(A) denote the primary outcome with treatment A. Thus, y(A) = 1 if a patient survives with observation and let y(A) = 0 if the patient does not survive. Let P denote the probability of a specified event. Then average patient welfare with observation is the survival probability P[y(A) = 1].

Let lymph node dissection be treatment B. Let y(B) = 1 if a patient survives with dissection and y(B) = 0 otherwise. Let s(B) denote the secondary outcome with treatment B. Thus, s(B) = 1 if a patient experiences

lymphedema and s(B) = 0 otherwise. Then average patient welfare with dissection is the difference between the survival probability and *h* times the probability of lymphedema; that is, P[y(B) = 1] − *h*·P[s(B) = 1]. The optimal treatment is the one yielding the higher average patient welfare. Thus, observation is optimal if P[y(A) = 1] exceeds P[y(B) = 1] − *h*·P[s(B) = 1]. Dissection is optimal if otherwise.

In this setting, a state of nature is a set of values for the four probabilities P[y(A) = 1], P[y(B) = 1], and P[s(B) = 1|y(B) = 0], and P[s(B) = 1|y(B) = 1]. Determination of the optimal choice is infeasible without knowledge of these probabilities. A trial such as MSLT-II yields information about the probabilities of survival and lymphedema that clinicians need to know to choose treatments maximizing average patient welfare. The sample size determines the extent of the information. For any positive constant ε, a sample of size N per treatment arm enables ε-optimal treatment if N is sufficiently large.

The Appendix explains the computation of sample sizes that enable ε-optimal treatment for any values of *h* and ε, assuming use of the empirical success rule. We show computational results using two methods to determine maximum regret, one applying simulated annealing and the other using a normal approximation to the finite-sample distribution of empirical success. Our computations of maximum regret are conservative in the sense that we impose no a priori restrictions on the values of the probabilities of survival and lymphedema. One may believe that some values of these probabilities are implausible. If so, one may restrict attention to the values deemed plausible and maximize regret across these values. Maximum regret across a restricted set of probability values logically must be less than or equal to maximum regret across all values. Hence, performing our analysis with restrictions imposed on the plausible probabilities cannot weaken our findings on minimum sample sizes that suffice to enable ε-optimal treatment choice. It may yield smaller minimum sample sizes than those we report in Section 3.

3. Findings

We report the minimal sample size enabling near-optimal treatment when $h = 0.2$ and $\varepsilon = 0.0085$. Recall that we use a scale for patient welfare in which, absent lymphedema, a patient has welfare 1 if he survives three years and welfare 0 if he dies within three years. Setting $h = 0.2$ supposes that suffering from lymphedema reduces a patient's welfare by one-fifth relative to lymphedema-free survival. This quantification of the welfare loss produced by lymphedema is suggested by Cheville *et al*. (2010), who elicited from a group of patients their perspectives on the matter. See Basu and Meltzer (2007) for discussion of elicitation of patient treatment and health preferences more generally.

Setting $\varepsilon = 0.085$ follows naturally from how the MSLT-II investigators performed their power calculation. They judged a difference of 5 percentage points in melanoma-specific survival to be a clinically meaningful loss in patient welfare and they judged 0.17 to be an acceptable probability of Type II error. As discussed in Section 2, regret equals the magnitude of welfare loss times the probability that the loss will occur. Thus, the MSLT-II investigators judged $0.17 \times 0.05 = 0.0085$ to be an acceptable level of regret.

When $h = 0.2$ and $\varepsilon = 0.0085$, we find that near-optimal treatment is achievable if one assigns 244 patients to observation and 244 to dissection. This total sample size of 488 is much smaller than the 1939 subjects enrolled in MSLT-II.

4. Discussion

Our reconsideration of MSLT-II opens many possibilities that the investigators could have contemplated if they had approached trial design from the perspective of near-optimal treatment rather than statistical

power. They could have achieved the declared study objective---comparison of observation and dissection for patients with a malignant sentinel lymph node---with a much smaller total sample size. Reducing total sample size would have lowered the cost of executing the trial, the time required to recruit subjects, and the need to perform trials across multiple centers. Or, maintaining enrollment of 1939 subjects, they could have expanded the study objective by performing a trial with more than two treatment arms, thus yielding information about more treatment options.

Our suggestion of potential alternatives to the MSLT-II design should not be interpreted as criticism of its investigators. They adhered to what has been the standard practice in setting sample size, proceeding as have thousands of other clinical trials. The lessons of Manski and Tetenov (2016) and this paper are intended to be forward looking, as trials are designed henceforth. The ideas developed in these papers have general potential application.

A coherent alternative to setting trial size to enable near-optimal treatment is application of Bayesian statistical decision theory. The Bayesian perspective would be attractive if trial designers and clinicians were able to place a credible consensus subjective prior distribution on treatment response. However, Bayesian statisticians have long struggled to provide guidance on specification of priors and the matter continues to be controversial. In the context of clinical trials, the authors and discussants of Spiegelhalter, Freedman, and Parmar (1994) express a spectrum of views. Bayesian decision theorists have occasionally recognized that inability to specify a credible subjective distribution may yield poor decisions. Berger (1985) cautions that (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.”

Appendix: Computation of Sample Sizes that Enable ε-optimal Treatment

A.1. Setup and Notation

In the setting described in the paper, the vector of unknown parameters is $\theta = (a, b_{00}, b_{01}, b_{10}, b_{11})$. The probability that treatment A yields a positive outcome (survival) is $a \equiv P_\theta(Y(A) = 1)$ and the probability of a negative outcome (death) is $1 - a$. Treatment B has four possible outcomes, with two values for the main outcome $Y(B) \in \{0,1\}$ and two for the side effect $S(B) \in \{0,1\}$. The probability of each outcome is $b_{ys} \equiv P_\theta(Y(B) = y, S(B) = s), y \in \{0,1\}, s \in \{0,1\}$. The parameter space for θ is $\Theta = [0,1]^5$, with the restriction that $b_{00} + b_{01} + b_{10} + b_{11} = 1$. This specification of the parameter space imposes no assumptions on the joint distribution of primary outcomes and side effects.

The patient's expected gain/loss from treatment B compared to treatment A in state of nature θ equals

$$\tau_\theta = \mathbb{E}_\theta\big(Y(B) - hS(B)\big) - \mathbb{E}_\theta\big(Y(A)\big).$$

With N subjects per treatment arm, a sample analog of this quantity computed from the trial data is

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(B) - hS_i(B)\big) - \frac{1}{N}\sum_{i=N+1}^{2N} Y_i(A)\,,$$

where observations of individuals randomly assigned to treatment B have indices $i = 1 \ldots N$ and those assigned to treatment A have indices $i = N + 1 \ldots 2N$.

We assume that the patient will choose between the treatment options using the *Empirical Success (ES)* rule, choosing treatment B if $\hat{\tau} > 0$, choosing treatment A if $\hat{\tau} < 0$, and choosing either treatment with probability ½ if $\hat{\tau} = 0$.

The regret of the ES rule in state θ is the product of the magnitude of error $|\tau_\theta|$ from choosing the suboptimal treatment and the probability of making that error:

$$R_N(\theta) = \begin{cases} -\tau_\theta[P_\theta(\hat{\tau} > 0) + \tfrac{1}{2}\, P_\theta(\hat{\tau} = 0)], & \tau_\theta \leq 0, \\ \tau_\theta[P_\theta(\hat{\tau} < 0) + \tfrac{1}{2}\, P_\theta(\hat{\tau} = 0)], & \tau_\theta > 0. \end{cases}$$

To determine whether a particular sample size N is sufficient to enable ε-optimal treatment choice when using the ES treatment rule, we need to compute the maximum regret of the ES rule $\max_{\theta \in \Theta} R_N(\theta)$ and compare it to ε. Exact computation of $\max_{\theta \in \Theta} R_N(\theta)$ is challenging because the function $R_N(\theta)$ is not convex in θ. It generally has multiple local maxima, both over $\{\theta: \tau_\theta < 0\}$ and over $\{\theta: \tau_\theta > 0\}$.

We consider two approximations for $\max_{\theta \in \Theta} R_N(\theta)$. The first method maximizes regret with an asymptotic normal approximation to $P_\theta(\hat{\tau} < 0)$. The derivations in the following section of the Appendix show that maximizing regret with this normal approximation over θ reduces to a very simple one-dimensional numerical optimization problem. The second method computes exact values of $R_N(\theta)$ for given values of θ and searches for the maximum using the simulated annealing algorithm, described in detail in Press *et al.* (2007). This method is often used to search numerically for the global maximum of a function that may have multiple local maxima. It is computationally intensive and is not guaranteed to find the exact maximum, but its precision can be improved by increasing the computation time. With sufficient computation time this method is more accurate, since the normal approximation clearly underestimates maximum regret, especially in small samples.

The maximum regret of the ES rule approximated using both methods is reported in Table 1 for sample sizes ranging from N = 10 to N = 250 and for values of the side effect disutility ranging from $h = 0$ to $h = 0.5$. A table with calculations of maximum regret for other values of N and h is available from the authors.

Comparison of the results obtained using the two methods shows that it is important to compute regret exactly in very small sample sizes, where the normal approximation tends to substantially underestimate maximum regret. For sample sizes N > 100, both methods produce similar results. It is then more practical to use the normal approximation as it is computationally much simpler.

For $h = 0.2$, the maximum regret of the ES rule is greater than 0.0085 for all sample sizes up to $N \leq 243$ and is smaller than 0.0085 for all sample sizes $N \geq 244$. Hence, assigning 244 subjects to each treatment is sufficient for near-optimal treatment choice with $\varepsilon = 0.0085$.

A.2. Details for the Normal Approximation

A simple analytical approximation to maximum regret can be obtained with a Normal approximation to the finite-sample distribution of $\hat{\tau}$. In a trial with N randomly drawn subjects per treatment arm,

$$\hat{\tau} \sim^{a} \mathcal{N}(\tau_\theta, V_\theta/N),$$

$$V_\theta = \mathrm{Var}_\theta\big(Y(B) - hS(B)\big) + \mathrm{Var}_\theta\big(Y(A)\big).$$

The regret of the ES rule in state θ equals

$$R_N(\theta) = \begin{cases} -\tau_\theta P_\theta(\hat{\tau} > 0), & \tau_\theta \leq 0, \\ \tau_\theta P_\theta(\hat{\tau} \leq 0), & \tau_\theta > 0, \end{cases}$$

which is approximated by

$$R_N^a(\theta) = \begin{cases} -\tau_\theta \Phi\big(\tau_\theta \sqrt{N/V_\theta}\big), & \tau_\theta \leq 0, \\ \tau_\theta \Phi\big(-\tau_\theta \sqrt{N/V_\theta}\big), & \tau_\theta > 0. \end{cases} \quad (A.1)$$

Note that for a given value of $\tau_\theta \leq 0$, $\Phi\big(\tau_\theta \sqrt{N/V_\theta}\big)$ is increasing in V_θ, hence $R_N^a(\theta)$ is increasing in V_θ. For a given value of $\tau_\theta > 0$, $\Phi\big(-\tau_\theta \sqrt{N/V_\theta}\big)$ and $R_N^a(\theta)$ are also increasing in V_θ. To find the maximum of $R_N^a(\theta)$ we will focus on a simple subset of parameter values that maximize V_θ for any τ_θ.

For any parameter vector $\theta = (a, b_{00}, b_{01}, b_{10}, b_{11})$, there exists an alternative parameter vector θ^* with the same expected gain/loss $\tau_\theta^* = \tau_\theta$, but with higher variance of the estimate $V_\theta^* \geq V_\theta$. We construct θ^* by spreading the probability mass b_{00} (on outcome $Y(B) - hS(B) = 0$) between b_{01} ($Y(B) - hS(B) = -h$) and b_{10} ($Y(B) - hS(B) = 1$). Moving $\frac{1}{1+h} b_{00}$ to b_{01} and $\frac{h}{1+h} b_{00}$ to b_{10} preserves the mean of $Y(B) -$

$hS(B)$, and hence τ_θ. Similarly, spreading the probability mass b_{11} (on outcome $Y(B) - hS(B) = 1 - h$) by moving $\frac{h}{1+h} b_{11}$ from b_{11} to b_{01} and by moving $\frac{1}{1+h} b_{11}$ from b_{11} to b_{10} also preserves the mean of $Y(B) - hS(B)$. In summary, $\theta^* = (a, 0, b_{01}^*, b_{10}^*, 0)$, where

$$b_{01}^* = b_{01} + \frac{1}{1+h} b_{00} + \frac{h}{1+h} b_{11},$$

$$b_{10}^* = b_{10} + \frac{h}{1+h} b_{00} + \frac{1}{1+h} b_{11}.$$

Since $\mathbb{E}_{\theta^*}\big(Y(B) - hS(B)\big) = \mathbb{E}_\theta\big(Y(B) - hS(B)\big)$, in order to show that $\mathrm{Var}_{\theta^*}\big(Y(B) - hS(B)\big) \geq \mathrm{Var}_\theta\big(Y(B) - hS(B)\big)$ it is sufficient to show that $\mathbb{E}_{\theta^*}\left(\big(Y(B) - hS(B)\big)^2\right) \geq \mathbb{E}_\theta\left(\big(Y(B) - hS(B)\big)^2\right)$.

$$\mathbb{E}_\theta\left(\big(Y(B) - hS(B)\big)^2\right) = h^2 b_{01} + b_{10} + (1-h)^2 b_{11},$$

$$\mathbb{E}_{\theta^*}\left(\big(Y(B) - hS(B)\big)^2\right) = h^2\left(b_{01} + \frac{1}{1+h} b_{00} + \frac{h}{1+h} b_{11}\right) + \left(b_{10} + \frac{h}{1+h} b_{00} + \frac{1}{1+h} b_{11}\right)$$

$$= h^2 b_{01} + b_{10} + \frac{h^2+h}{1+h} b_{00} + \frac{h^3+1}{1+h} b_{11}.$$

Since $\frac{h^3+1}{1+h} = 1 - h + h^2 > 1 - 2h + h^2 = (1-h)^2$ for $h > 0$, it follows that $\frac{h^3+1}{1+h} b_{11} \geq (1-h)^2 b_{11}$. Hence, $\mathbb{E}_{\theta^*}\left(\big(Y(B) - hS(B)\big)^2\right) \geq \mathbb{E}_\theta\left(\big(Y(B) - hS(B)\big)^2\right)$.

Under θ^*, the outcome values $Y(B) - hS(B)$ of treatment B have a binary distribution on $\{-h, 1\}$, with probabilities $b_{01}^* = 1 - b_{10}^*$ and b_{10}^*. Hence,

$$\mathrm{Var}_{\theta^*}\big(Y(B) - hS(B)\big) = (1+h)^2 b_{10}^*(1 - b_{10}^*).$$

and

$$V_{\theta^*} = (1+h)^2 b_{10}^*(1 - b_{10}^*) + a(1-a), \qquad (A.2)$$

$$\tau_\theta^* = b_{10}^* - h(1 - b_{10}^*) - a. \qquad (A.3)$$

Thus

$$\max_{\theta\in\Theta} R_N^a(\theta) = \max_{\theta\in\Theta^*} R_N^a(\theta),$$

where $\Theta^* = \{\theta \in \Theta \text{ s.t. } \theta = (a, 0, 1 - b_{10}^*, b_{10}^*, 0), a \in [0,1], b_{10}^* \in [0,1]\}$.

We can further constrain the set of values of (a, b_{10}^*) on which V_{θ^*} may attain its maximum. The value of the treatment effect $\tau_\theta^*(a, b_{10}^*)$ as a function of $\begin{bmatrix} a \\ b_{10}^* \end{bmatrix}$ remains constant along the vector $\mathbf{u} = \begin{bmatrix} 1+h \\ 1 \end{bmatrix}$. The directional derivative of $V_{\theta^*}(a, b_{10}^*)$ in direction $\mathbf{u}$ equals

$$\nabla_{\mathbf{u}} V_{\theta^*} = \nabla V_{\theta^*} \cdot \mathbf{u} = [(1-2a) \quad (1+h)^2(1-2b_{10}^*)] \cdot \begin{bmatrix} 1+h \\ 1 \end{bmatrix} =$$

$$= (1+h)\big((1-2a) + (1+h)(1-2b_{10}^*)\big).$$

Hence, over the set of (a, b_{10}^*) values on which τ_θ^* is constant, V_{θ^*} is increasing in direction $\mathbf{u}$ when

$$(1-2a) + (1+h)(1-2b_{10}^*) > 0$$

(for low values of a and b_{10}^*) and reaches its maximum at the point where

$$(1-2a) + (1+h)(1-2b_{10}^*) = 0,$$

i.e.,

$$b_{10}^* = \frac{1 + h/2 - a}{1+h}$$

if this point is feasible or at the closest feasible point. It follows that the set of parameter values (a, b_{10}^*) which maximize variance V_{θ^*} for different values of τ_θ^* consists of three line segments:

$$a = 0,\ b_{10}^* \in \left[\frac{1+h/2}{1+h}, 1\right], \tag{A.4}$$

$$a \in [0,1], b_{10}^* = \frac{1 + h/2 - a}{1+h}, \tag{A.5}$$

$$a = 1, b_{10}^* \in \left[0, \frac{h/2}{1+h}\right]. \tag{A.6}$$

References

Basu, A. and D. Meltzer (2007), "Value of information on preference heterogeneity and individualized care," *Medical Decision Making*, 27, 112-27.

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

Cheville, A., M. Almoza, J. Courmier, and J. Basford (2010), "A Prospective Cohort Study Defining Utilities Using Time Trade-Offs and the Euroqol-5D to Assess the Impact of Cancer-Related Lymphedema," *Cancer*, 116, 3722-3731.

Faries, M. (2018), "Completing the Dissection in Melanoma: Increasing Decision Precision," *Annals of Surgical Oncology*, https://doi.org/10.1245/s10434-017-6330-4.

Faries, M. *et al.* (2017), "Completion Dissection or Surveillance for Sentinel-Node Metastasis in Melanoma," *New England Journal of Medicine,* 376, 2211-2222.

Halpern S, Karlawish J, Berlin J (2002), "The Continued Unethical Conduct of Underpowered Clinical Trials," *Journal of the American Medical Association,* 288, 358-362.

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.

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

Manski, C. (2018), "Treatment Choice with Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing," *The American Statistician*, forthcoming.

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

Meltzer D. (2001), "Addressing Uncertainty in Medical Cost-Effectiveness: Implications of Expected Utility Maximization for Methods to Perform Sensitivity Analysis and the Use Of Cost-Effectiveness Analysis to Set Priorities for Medical Research," *Journal of Health Economics,* 20, 109-129.

Phelps C. and A. Mushlin (1988), "Focusing Technology Assessment using Medical Decision Theory," *Medical Decision Making,* 8, 279-289.

Press, W., S. Teukolsky, W. Vetterling, and B. Flannery (2007), *Numerical Recipes: The Art of Scientific Computing, Third Edition*, Cambridge: Cambridge University Press

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

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

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

Tetenov, A. (2016), "An Economic Theory of Statistical Testing," Cemmap working paper CWP50/16, doi: 10.1920/wp.cem.2016.5016.

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

Table 1: Near-optimality (maximum regret) of ES rules
(N is the number of subjects per treatment arm)

		$h = 0$		$h = 0.1$		$h = 0.2$		$h = 0.3$		$h = 0.4$		$h = 0.5$	
N =		Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.	Simulated annealing	Normal approx.
10		0.038209	0.037490	0.044905	0.039672	0.045017	0.041857	0.045794	0.044046	0.046704	0.046237	0.049236	0.048431
20		0.026947	0.026689	0.030401	0.028180	0.030479	0.029672	0.031212	0.031166	0.032803	0.032661	0.034487	0.034157
30		0.021983	0.021841	0.024046	0.023039	0.024516	0.024237	0.025874	0.025435	0.026805	0.026634	0.028039	0.027834
40		0.019029	0.018937	0.020710	0.019963	0.021105	0.020989	0.021930	0.022016	0.023172	0.023044	0.024218	0.024071
50		0.017016	0.016949	0.018217	0.017860	0.018829	0.018772	0.019865	0.019683	0.020688	0.020595	0.021621	0.021507
60		0.015530	0.015480	0.016640	0.016307	0.017170	0.017134	0.018019	0.017962	0.018859	0.018789	0.019709	0.019617
70		0.014376	0.014336	0.015231	0.015099	0.015890	0.015861	0.016708	0.016624	0.017444	0.017387	0.018227	0.018150
80		0.013447	0.013414	0.014291	0.014124	0.014861	0.014835	0.015612	0.015546	0.016306	0.016257	0.017034	0.016968
90		0.012677	0.012649	0.013371	0.013317	0.014009	0.013985	0.014690	0.014653	0.015365	0.015321	0.016048	0.015990
100		0.012025	0.012002	0.012724	0.012634	0.013287	0.013266	0.013952	0.013898	0.014570	0.014530	0.015215	0.015163
150		0.009817	0.009804	0.010330	0.010316	0.010841	0.010827	0.011359	0.011339	0.011876	0.011850	0.012395	0.012362
200		0.008501	0.008493	0.008941	0.008933	0.009384	0.009374	0.009826	0.009814	0.010274	0.010255	0.010720	0.010696
250		0.007603	0.007597	0.007995	0.007990	0.008390	0.008382	0.008786	0.008775	0.009183	0.009168	0.009580	0.009560