Supplemental Materials for "Statistical Decision Properties of Imprecise Trials Assessing Coronavirus Disease 2019 (COVID-19) Drugs"

Charles F. Manski, PhD, Aleksey Tetenov, PhD

***Value in Health*, 2021**

https://doi.org/10.1016/j.jval.2020.11.019

Appendix A: Technical Details

General setup

We use concepts and notation like those in Manski[1] and Manski and Tetenov[2-3]. The clinician must assign one of L treatments studied in the clinical trial to each member of a treatment population, denoted J. Denote the set of treatments by $T = \{1, 2, \ldots, L\}$, treatment 1 being standard care. Each individual $j \in J$ has a response function $y_j(\cdot)$: $T \rightarrow Y$ mapping treatments $t \in T$ into individual patient-relevant outcomes $y_j(t) \in Y$. In general, outcomes could be multi-valued and multi-dimensional. For example, the relevant outcomes for COVID-19 treatment may be survival, taking the value 0 or 1, and time to recovery for those who survive, measured in number of days.

The probability distribution $P[y(\cdot)]$ of the random function $y(\cdot)$: $T \rightarrow Y$ describes treatment response across the population. The distribution P is unknown. The set of all feasible distributions P is $\{P_s, s \in S\}$, where S indexes all feasible *states of nature*. When computing near-optimality in Tables 2 and 4, we include in S all logically possible outcome distributions.

We assume that *patient welfare* is a known function u: $Y \rightarrow \mathbf{R}$ of individual outcomes. For binary outcomes $Y = \{0, 1\}$, with 1 denoting success. In this special case, it is without loss of generality to set $u(y) = y$. For two-dimensional patient outcomes $y = (y_p, y_{se})$, where y_p denotes the primary outcome and y_{se} the side effect, Manski and Tetenov[3] considered patient welfare that is a weighted sum of the two outcomes: $u(y) = y_p - hy_{se}$, where h expresses the harm of the side effect relative to the primary outcome.

Now consider data generation. Let Ψ denote the sample space; that is, Ψ is the set of data samples that could be generated by the trial. Let Q_s denote the sampling distribution on Ψ in state of nature s. That is, Q_s is the probability distribution of different trial outcomes.

We consider trials that randomize a predetermined number of subjects n_t to each treatment t. The set $n_T \equiv [n_t, t \in T]$ of stratum sample sizes defines the design. The total number of subjects in the trial is then $N \equiv \sum_{t \in T} n_t$. The data ψ are the N pairs of individual treatment assignments t_i and outcomes y_i: $\psi = [(t_i, y_i), i = 1, 2, \ldots, N]$.

The sampling distribution Q_s is determined by the probability distribution of treatment response P_s and the trial design, with $Q_s(y_i|t_i) = P_s(y(t_i))$. We assume that treatment response is individualistic; that is, patient outcomes are statistically independent of the outcomes of other patients in the trial.

A statistical treatment rule maps sample data into a treatment allocation. A feasible treatment rule is a function that randomly allocates persons across the different treatments. Let Δ now denote the space of functions that map T into the unit interval and that satisfy the adding-up condition: $\delta \in \Delta \Rightarrow \sum_{t \in T} \delta(t, \psi) = 1, \ \forall \psi \in \Psi$. Then each function $\delta \in \Delta$ defines a statistical treatment rule.

The mean welfare outcome of treatment t in state of nature s is denoted by $\mu_{st} \equiv E_s[u(y(t))]$. The maximum average patient welfare achievable in state s is $\max_{t \in T} \mu_{st}$. After trial data ψ are observed, the fraction $\delta(t, \psi)$ of patients will be treated with treatment t, resulting in mean patient welfare $\sum_{t \in T}\big(\mu_{st}\delta(t, \psi)\big)$. The mean welfare of patients treated according to statistical treatment rule δ over repeated realizations of the trial is then $\int_\Psi \sum_{t \in T}\big(\mu_{st}\delta(t, \psi)\big)\, dQ_s(\psi) = \sum_{t \in T} \mu_{st} E_s[\delta(t, \psi)]$, where $E_s[\delta(t, \psi)] = \int_\Psi \delta(t, \psi) dQ_s(\psi)$ is the expected (across potential samples) fraction of persons who will be assigned to treatment t.

Application of statistical treatment rule δ in state of nature s leads to an expected loss (regret) equal to

(A1) $\max_{t \in T} \mu_{st} - \sum_{t \in T} \mu_{st} E_s[\delta(t, \psi)]$.

The near-optimality (maximum regret) of statistical treatment rule δ is the maximum value of (A1) over all feasible states of nature:

$$\text{(A2)} \quad \max_{s \in S}\left(\max_{t\in T}\mu_{st} - \sum_{t\in T}\mu_{st}E_s[\delta(t,\psi)]\right).$$

Hypothesis Testing Rules

First, we consider statistical treatment rules based on hypothesis tests for univariate outcomes y. Denote the sample mean of y observed in arm t of the trial by $\bar{y}_t = \frac{1}{n_t}\sum_{i:t_i=t} y_i$. To test the null hypothesis that all treatments have the same outcome distribution, we use $\hat{\sigma}^2 = \frac{1}{N-L}\sum_{t\in T}\sum_{i:t_i=t}(y_i - \bar{y}_t)^2$ as the estimator of common variance. Then the t-statistic for comparing the mean outcome of treatment $t = 2,\ldots,L$ with that of standard care (treatment 1) equals $\tau_t = \frac{\bar{y}_t - \bar{y}_1}{\hat{\sigma}\sqrt{1/n_t + 1/n_1}}$. Let c be the critical value adjusted for multiplicity. Specifically, we use the Student's t-distribution for two-arm trials and the Dunnett's test critical value for multiple comparisons for multi-arm trials.

The hypothesis test rule prescribes treatment 1 (standard care) to everyone if all t-statistics are below the critical value.:

$$\delta_H(1,\psi) \equiv 1\left\{\max_{t\in\{2,\ldots,L\}}\tau_t \le c\right\}.$$

If some t-statistics comparing treatments $2,\ldots,L$ to standard care exceed the critical value, these treatments are considered statistically significantly better than standard care. We assume that among these treatments the one with the largest mean outcome in the trial will be prescribed, with equal probability if there is a tie.

$$\delta_H(t,\psi) \equiv \frac{1\left\{\tau_t > c,\ \bar{y}_t = \max_{t'\in\{2,\ldots,L\}}\bar{y}_{t'}\right\}}{\sum_{t'\in\{2,\ldots,L\}}1\left\{\tau_t > c,\ \bar{y}_t = \max_{t'\in\{2,\ldots,L\}}\bar{y}_{t'}\right\}}.$$

When treatment arms 2,…,L have equal sample sizes, as in our Table 4, the t-statistics τ_t have the same ranking as the sample means $\bar{y}_t$. Hence, prescribing the treatment with the largest mean outcome in the trial is equivalent in this case to prescribing the treatment with the largest t-statistic.

The Empirical Success Rule

Let $\bar{u}_t = \frac{1}{n_t}\sum_{i:t_i=t} u(y_i)$ denote the mean patient welfare observed in treatment arm $t = 1, 2, \ldots, L$. The empirical success rule considers all treatments in the trial symmetrically and prescribes the treatment with the largest observed mean patient welfare. If there is a tie, all treatments with the largest observed mean patient welfare are prescribed with equal probability.

$$\delta_{ES}(t,\psi) \equiv \frac{1\left\{\bar{u}_t = \max_{t' \in \{1,\ldots,L\}} \bar{u}_{t'}\right\}}{\sum_{t' \in \{1,\ldots,L\}} 1\left\{\bar{u}_t = \max_{t' \in \{1,\ldots,L\}} \bar{u}_{t'}\right\}}.$$

For binary outcomes, we take $u(y) = y$.

Computation of Near-Optimality for Two-Arm Trials with Binary Outcomes

When computing the near-optimality results reported in Table 2, we consider the set of all possible distributions of binary outcomes with means $p_1 \equiv \mathrm{E}[y(1)]$, $p_2 \equiv \mathrm{E}[y(2)]$, $(p_1, p_2) \in [0, 1]^2$.

Let m_1 and m_2 denote the number of positive outcomes in each arm of the trial. For binary outcomes, $\psi = (m_1, m_2)$ is a sufficient statistic for the sample. Hence, it is sufficient to consider the sample space $\Psi = \{0, 1, \ldots, n_1\} \times \{0, 1, \ldots, n_2\}$. The probability density function of ψ is a product of two binomial density functions. This sample space is sufficiently small, so we compute (A1) exactly.

The function (A1) is continuous in (p_1, p_2) but may have multiple global and local maxima. We approximate the maximum in (A2) by grid search using 1000 possible values for each parameter equally spaced on [0,1]: {0.0005, 0.0015, …, 0.9995}.

Computation of Near-Optimality for Multi-Arm Trials with Binary Outcomes

To compute the results reported in Table 4, we consider the set of all possible distributions of binary outcomes with means $p_t \equiv \mathrm{E}[y(t)]$, $t = 1, \ldots, L$, $(p_1, \ldots, p_L) \in [0, 1]^L$. Let m_t denote the number of positive outcomes in arm t of the trial. For binary outcomes, $\psi = (m_1, \ldots, m_L)$ is a sufficient statistic for the sample. Hence, we consider the sample space $\Psi = \{0, 1, \ldots, n_1\} \times \ldots \times \{0, 1, \ldots, n_L\}$. The large size of the sample space makes it impractical to evaluate (A1) exactly. Instead, given each value of $(p_1, \ldots, p_L)$ we simulate a large number of trial outcomes to approximate the sampling distribution Q_s. Our computations of the maximum of (A2) proceed in three steps.

First, we conduct a grid search using 51 possible values for each parameter $p_t \in [0, 0.02, \ldots, 1]$. For each combination of parameters, we approximate the sampling distribution Q_s by simulating 100,000 trial outcomes. The results of this grid search suggest that the largest expected loss for the empirical success rule occurs when the parameters have the form $p_1 = a$, $p_2 = p_3 = p_4 = p_5 = b$, $a > b$. The largest expected loss for the Dunnett's test rule occurs when $p_1 = a$, $p_2 = b$, $p_3 = p_4 = p_5 = c$, $b > a$, $b > c$.

In the second step, we conduct a grid search over these two lower-dimensional parameter spaces using 101 possible parameter values from $[0, 0.01, \ldots, 1]$ for a, b, and c. In this step we approximate Q_s by simulating 1,000,000 trial outcomes.

In the last step, we take 10 parameter combinations yielding the largest estimated expected loss for each decision rule in step 2 and re-compute expected loss by simulating 100,000,000 trial outcomes. We do this to verify that our results are not affected by bias resulting from approximating Q_s by simulation.

The MATLAB code used to perform the computations is available in Supplemental Materials at https://doi.org/10.1016/j.jval.2020.11.019.

Appendix B: Summary of Findings on Near-Optimality of the Empirical Success Rule with Patient-Specific Treatment and Multiple Outcomes

Near-Optimality with Binary Primary and Secondary Outcomes

Manski and Tetenov[3] study the near-optimality of the empirical success rule when there are two feasible treatments and patient welfare is a weighted sum of binary primary and secondary outcomes. The primary outcome is patient survival for a specified time period. The secondary one denotes whether the patient suffers a specified side effect of treatment.

When a patient does not suffer the side effect, we let welfare equal 1 if a patient survives and equal 0 if he does not survive. When a patient experiences the side effect, welfare is lowered by a specified fraction h, whose value expresses the harm associated with the side effect. Thus, a patient who experiences the side effect has welfare $1 - h$ if he survives and $-h$ if he does not survive.

Manski and Tetenov[3] develop an algorithm to compute the near-optimality of the empirical success rule, which evaluates trial data by the frequencies of survival and the side effect observed with each treatment. We present numerical findings for alternative values of sample size and the value h expressing the harm of the side effect.

Near-Optimality with Bounded Outcomes

Exact computation of near-optimality is feasible when trial outcomes are binary or take only a few values, but it becomes more onerous when outcomes can take many values or are continuous. When outcomes are bounded, large-deviations inequalities of probability theory yield upper bounds on the near-optimality of the empirical success rule. These upper bounds provide conservative measures of near-optimality. Their value is that they are simple to compute and are sufficiently informative to provide useful guidance to clinicians.

Research of this type was initiated by Manski[1], who used a large-deviations inequality for sample averages of bounded outcomes to derive an upper bound on the near-optimality of the empirical success rule when used to choose between two treatments. Manski and Tetenov[2] extended the analysis to multi-

arm trials. Their Proposition 1 extends the early finding of Manski[1] from two to multiple treatments. Proposition 2 derives a new large-deviations bound for multiple treatments.

Let L be the number of treatment arms and let V be the range of the bounded outcome. When the trial has a balanced design, with n subjects per treatment arm, the upper bounds on near-optimality proved in Propositions 1 and 2 have simple forms, being $(2e)^{-½}V(L-1)n^{-½}$ and $V(\ln L)^{½}n^{-½}$. The former result provides a tighter bound than the latter for two or three treatments, while the latter result gives a tighter bound for four or more treatments. In both cases, the upper bound decreases toward zero at rate √n as the number n of subjects per arm increases.

Near-Optimality with Heterogeneous Patients

Patient response to treatments for COVID-19 may be heterogenous, varying with covariates including age, gender, and comorbidities. Hence, a clinician may want to assess the near-optimality of a decision criterion when applied to patients who share similar observed covariates.

In principle, this is easy to do. The clinician may view each group of patients who share similar covariates as a separate patient population. Accordingly, the clinician may apply the empirical success rule separately to each group, choosing a treatment that yields the highest average outcome among the trial participants who have the group covariates. In this manner, patient care may recognize heterogeneity of treatment response. See Manski[1] and Manski and Tetenov[3].

In practice, the ability of clinicians to differentially treat patients with different covariates is sometimes limited by the failure of medical researchers to report how trial findings vary with patient covariates. A common rationale is concern with statistical significance. Stratifying trial participants into covariate groups usually reduces the statistical precision of estimates of treatment effects. Research articles often report only findings that are statistically significant by conventional criteria.

Information is lost when reporting research findings is tied to statistical significance. It is important to study and report observable heterogeneity in treatment response to the extent feasible. The analysis of

this paper makes clear that estimates of treatment effects need not be statistically significant to be clinically useful.

References

1. Manski C. Statistical treatment rules for heterogeneous populations. *Econometrica*. 2004;72(4):221–246.

2. Manski C, Tetenov A. Sufficient trial size to inform clinical practice. *Proc Natl Acad Sci*. 2016;113(38):10518–10523.

3. Manski C, Tetenov A. Trial size for near-optimal treatment: reconsidering MSLT-II. *Amer Stat*. 2019;73(S1):305–311.