Statistical Decision Properties of Imprecise Trials Assessing COVID-19 Drugs

Charles F. Manski and Aleksey Tetenov (MedRxiv, 2020)

Background

- "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica* 72, 2004, 221-246.
- "Sufficient Trial Size to Inform Clinical Practice," *Proceedings of the National Academy of Sciences* 113, 2016, 10518-10523.
- "Trial Size for Near-Optimal Choice Between Surveillance and Aggressive Treatment: Reconsidering MSLT-II," *The American Statistician* 73:sup1, 2019, 305-311.

Example of current practice

- Cao et al., “A Trial of Lopinavir–Ritonavir in Adults Hospitalized with Severe Covid-19,” *NEJM, 18 March 2020.*

- RCT
  - 99 patients assigned to receive lopinavir–ritonavir + ”standard care”
  - 100 patients assigned to “standard care” alone
  - Measured outcomes up to 28 days after randomization

Reported trial outcomes

- Primary Finding: “In a modified intention-to-treat analysis, lopinavir–ritonavir led to a median time to clinical improvement that was shorter by 1 day than that observed with standard care (hazard ratio, 1.39; 95% CI, 1.00 to 1.91).”

- Secondary Finding: “Mortality at 28 days was similar in the lopinavir–ritonavir group and the standard-care group (19.2% vs. 25.0%; difference, −5.8 percentage points; 95% CI, −17.3 to 5.7).”

Conclusions from the trial

- Cao et al. “no benefit was observed with lopinavir–ritonavir treatment beyond standard care.”

- U.S. NIH panel guidelines then recommended against the use of lopinavir/ritonavir writing: “lopinavir/ritonavir was studied in a small randomized controlled trial in patients with COVID-19 with negative results.”
    - This trial was the main piece of evidence, summarized as: “No difference in primary outcome (time to clinical improvement) was observed, and 28-day mortality was similar between groups.”

Questions

- How should we measure precision of an RCT?
  - Maximum expected loss in patient welfare for treatment chosen based on an RCT relative to the unknown best treatment. (*maximum regret*)
  - This depends on how the trial results are translated into clinical decisions. (*statistical treatment rule*)

- How should we use the results of clinical trials to decide which treatment to use?
  - Prevailing practice is to use a two-sided 5% hypothesis test to reach a binary conclusion: Standard care if the null isn’t rejected; innovation if the null is rejected with a significant positive estimate of average treatment effect.
  - We argue for the **Empirical Success** rule: choose the treatment with better average outcome and measure the outcome that patients want to maximize!

What happens in a trial with 100:99 patients using 28-day mortality as the outcome?

- Let mortality rate with standard care = 0.25 and use the standard t-test rule:

	Mortality rate with new treatment				
	0.35	0.30	0.25	0.20	0.15
% of trials after which standard care will be prescribed:	99.98%	99.7%	97.5%	86.76%	57.36%
Loss from choosing standard care:	0	0	0	0.05	0.10
% of trials after which new treatment will be prescribed:	0.02%	0.3%	2.5%	13.24%	42.64%
Loss from choosing treatment:	0.10	0.05	0	0	0
Expected loss:	0.0000	0.0002	0	0.0434	0.0574

- Maximum expected loss occurs when the new treatment has mortality rate 0.548 and standard care has rate 0.661. Then expected loss is (0.661 – 0.548) x error probability 0.624 = 0.071.

- Same scenarios, using the empirical success rule

	Mortality rate with new treatment				
	0.35	0.30	0.25	0.20	0.15
% of trials after which standard care will be prescribed:	94.28%	79.61%	51.64%	21.18%	4.22%
Loss from choosing standard care:	0	0	0	0.05	0.10
% of trials after which treatment will be prescribed:	5.72%	20.39%	48.36%	78.82%	95.78%
Loss from choosing treatment:	0.10	0.05	0	0	0
Expected loss:	0.0057	0.0102	0	0.0106	0.0042

- Maximum expected loss occurs when the new treatment has mortality rate 0.527 and standard care has rate 0.473. Then expected loss is (0.527 − 0.473) x error probability 0.226 = 0.012. The same expected loss occurs when standard care has mortality 0.527 and the new treatment 0.473.

Why use the empirical success rule?

- Theoretical study proves that it exactly or approximately minimizes maximum expected loss
  - Exactly optimal in balanced trials with binary outcomes (Stoye, *JoE*, 2009)
  - Asymptotically optimal in other two-arm trials (Hirano & Porter, *ECMA*, 2009)
- Treats Type I and Type II errors symmetrically
- Hypothesis testing treats the two errors asymmetrically. Maximum loss when the innovation is better is 250 times greater than maximum loss when it is worse.

Why we shouldn’t treat the two options asymmetrically

- “Standard care” for COVID-19 has been postulated without evidence than it is better than other options.
- If we were to start with a different definition of standard care, we would be stuck with it for a long time.
- Clinical equipoise
    - EMA: “There should be equipoise (“genuine uncertainty within the expert medical community [...] about the preferred treatment”) at the beginning of a randomised trial.”
    - One might motivate asymmetric decision-making after trials by having asymmetric Bayesian priors,
    - but interpreting ethical guidelines for starting trials through a Bayesian lens suggests that experts must
        1. Have disagreeing priors
        2. Some priors must favor one treatment, some the other

Multiple outcomes (side effects)

- Hypothesis testing does not protect against side effects outweighing benefits in primary outcome:
  - In sufficiently large trials, even small differences in “primary outcomes” will be detected, leading to headline conclusions that a new therapy is “effective”
  - Researcher definitions of primary outcomes often differ from patient-relevant outcomes (e.g., mortality)
- Empirical success rule can be applied to weighted averages of **all** patient-relevant outcomes observed in the trial
  - provided that patient-relevant outcomes are reported.

What sample sizes are sufficient?

- For two-armed trials with binary outcomes, using the empirical success rule yields these maximum expected losses:

Sample sizes	Near-optimality
20:20	0.0269
50:50	0.017
100:100	0.012
200:200	0.0085
500:500	0.0054
1000:1000	0.0038
4000:4000	0.0019
15000:15000	0.001

Downside of large sample sizes required by conventional testing rules

- Delay: It takes longer to recruit patients; hence, longer to reach conclusions.
- Crowds out trials of other treatments.
- Statistical significance requirement impedes subgroup analyses
    - There may be substantial heterogeneity in treatment effectiveness and the prevalence of side effects (e.g., by age)
    - The welfare weights attached to different outcome measures may vary with patient attributes.

Clinical trial landscape

- There are many alternative treatments in trials now
    - Each trial has a different set of inclusion criteria, a different PI, and only tests 1 innovation against its own definition of standard care.
    - Study populations differ across trials.
- It will be difficult to compare alternative treatments across trials.

Mutli-arm trials for Covid-19

- UK nationwide ”Recovery” trial started with 5 arms
  - Standard care
  - Lopinavir-Ritonavir
  - Low dose corticosteroids (dexamethasone)
  - Hydroxychloroquine
  - Azithromycin
- Patients were assigned to treatments in a 2:1:1:1:1 ratio

- WHO organized an international “Solidarity” trial with 5 arms
  - Standard care
  - Remdesivir
  - Lopinavir-Ritonavir
  - Lopinavir-Ritonavir plus Interferon beta-1a
  - Chloroquine or hydroxychloroquine
- These trials allow comparisons of multiple treatments on same population.

Evaluating multi-arm trials such as Recovery

- The Recovery protocol calls for results to be analyzed using Dunnett’s test. This is a multiple t-test procedure, with 0.05 Type I error probability that at least one test yields a positive statistically significant ATE. Presumably, the innovation with highest average outcome will be selected among those that pass the significance test. Otherwise, standard care will be selected.

- We contrast this with the empirical success rule, which selects the treatment with the highest average outcome, regardless of statistical significance.

- In practice, 3 treatment arms were stopped at different times. The results for each treatment were analyzed separately as if coming from a two-arm trial.

	Standard care	A	B	C	D
Sample size in each arm	500	250	250	250	250
Mortality rate of each treatment	0.25	0.15	0.20	0.30	0.35
Panel A: What happens if treatment decisions are made using two-sided Dunnett's test at 5% significance					
% of trials after which new treatment will be prescribed	25.65%	70.60%	3.75%	0	0
Loss from prescribing each treatment	0.1	0	0.05	0.15	0.2
Probability of error times the magnitude of loss	0.0257	0	0.0019	0	0
Expected loss given these mortality rates					0.0275
Panel B: What happens if treatment decisions are made using the empirical success rule					
% of trials after which new treatment will be prescribed	0.02%	92.95%	7.03%	0	0
Loss from prescribing each treatment	0.1	0	0.05	0.15	0.2
Probability of error times the magnitude of loss	0	0	0.0035	0	0
Expected loss given these mortality rates					0.0035

What sample sizes are sufficient?

- For five-armed trials with binary outcomes and 2:1:1:1:1 sample ratio, choosing the treatment using the empirical success rule and Dunnett's test rule imply the following maximum expected losses:

Sample sizes per arm	Near-optimality using Empirical Success rule	Near-optimality using Dunnett's test rule
100:50:50:50:50	0.0362	0.1224
200:100:100:100:100	0.0256	0.0855
500:250:250:250:250	0.0160	0.0532
1000:500:500:500:500	0.0112	0.0380
2000:1000:1000:1000:1000	0.0080	0.0274

- It is slightly better to divide the sample into equal-sized arms for the ES rule.

Near-optimality of empirical success rule with patient-specific treatment and multiple outcomes

- The above calculations concern settings where patients are observationally identical and trial outcomes are binary.

- In clinical practice, trial outcomes may take multiple values. Trials of COVID-19 drugs may report mortality outcomes and time to recovery for patients who survive. Patients may vary in treatment response by age, gender, and comorbidities.

- Methodological research has shown how to compute or bound the near-optimality of the empirical success rule when applied in a broad range of settings.

Near-optimality with binary primary and secondary outcomes

- Manski and Tetenov (2019) study near-optimality of the empirical success rule when there are two treatments and patient welfare is a weighted sum of binary primary and secondary outcomes. The primary outcome is survival. The secondary one denotes whether the patient suffers a specified side effect.

- 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*. Thus, a patient with the side effect has welfare 1 − *h* if he survives and –*h* if he does not survive.

- We develop an algorithm to compute the near-optimality of the empirical success rule.

Near-optimality with bounded outcomes

- Exact computation of near-optimality becomes onerous when outcomes can take many discrete values or are continuous.

- When outcomes are bounded, large-deviations inequalities yield upper bounds on the near-optimality of the empirical success rule. These bounds are simple to compute and are sufficiently informative to provide useful guidance to clinicians.

- Manski (2004) used the Hoeffding inequality for sample averages to derive an upper bound on near-optimality when there are two treatments.

- Manski and Tetenov (2016) extended the analysis to multi-arm trials. Let L be the number of treatments and V be the range of the outcome. When the trial has a balanced design, with n subjects per arm, the upper bounds on near-optimality are $(2\mathrm{e})^{-½}V(L-1)n^{-½}$ and $V(\ln L)^{½}n^{-½}$. The former is tighter than the latter for two or three treatments. The latter is tighter for four or more treatments.

Near-optimality with heterogeneous patients

- Patient response to treatments may vary with observed covariates. A clinician can assess the near-optimality of a decision criterion when applied to patients with similar covariates.

- In principle, a clinician may view each group of patients with similar covariates as a separate population and may apply the empirical success rule separately to each group.

- In practice, the ability to differentially treat patients with different covariates is limited by the failure of medical researchers to report how trial findings vary with covariates. Research articles often report only subgroup findings that are statistically significant.

- Information is lost when reporting research findings is tied to statistical significance. The analysis of this paper makes clear that estimates of treatment effects need not be statistically significant to be clinically useful.

Topics for future research

- We have considered treatment choice using data from one trial with full validity.

- Internal validity may be compromised by non-compliance and loss to follow up. External validity may be compromised by measurement of surrogate outcomes and study of patients who differ from those that clinicians treat in practice. The concept of near-optimality is applicable when analyzing data from trials with limited validity, but the calculations made in this paper require modification.

- A clinician may learn the findings of multiple trials and may be informed by observational data. Near-optimality is well-defined in these settings, but methods for application are yet to be developed.

- A further issue concerns dynamic treatment choice when new trials and observational data may emerge in the future. The concept of near-optimality should be extendable, but methodology is yet to be developed.
  - Dynamic analysis of treatment choice made with hypothesis tests may be especially difficult, because testing views standard care and new treatments asymmetrically. As new data accumulate, the designation of standard care may change, leading to a change in the null hypothesis when new trials are evaluated.

Technical Appendix

We use concepts and notation in Manski (2004) and Manski and Tetenov (2016, 2019).

The clinician must assign one of L treatments studied in the trial to each member of treatment population J.

Denote treatments by $T = \{1, 2, \ldots, L\}$, with t = 1 being standard care.

Each $j \in J$ has a response function $y_j(\cdot)$: $T \rightarrow Y$ mapping treatments $t \in T$ into patient-relevant outcomes $y_j(t) \in Y$. Outcomes can be multi-valued and multi-dimensional. Treatment response is individualistic.

The distribution $P[y(\cdot)]$ of the random function $y(\cdot)$: $T \rightarrow Y$ describes treatment response across the population. The set of feasible distributions is $\{P_s, s \in S\}$, S indexing feasible *states of nature*.

In Tables 2 and 4, we include in S all logically possible outcome distributions.

Patient welfare is a known function u: $Y \rightarrow \mathbf{R}$ of individual outcomes.

For binary outcomes $Y = \{0, 1\}$, with 1 denoting success, and $u(y) = y$. For two-dimensional outcomes $y = (y_p, y_{se})$, where y_p denotes the primary outcome and y_{se} a side effect, Manski and Tetenov (2019) considered welfare function $u(y) = y_p - hy_{se}$.

Consider data generation. Ψ denotes the sample space. Q_s denotes the sampling distribution on Ψ in state of nature s. Q_s is the distribution of 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 sample sizes defines the design. The total number of subjects is $N \equiv \sum_{t \in T} n_t$. The data ψ are the N pairs of individual treatment assignments t_i and outcomes y_i: $\psi = [(t_i, y_i), i = 1, 2, \ldots, N]$.

Q_s is determined by the distribution of treatment response P_s and the trial design, with $Q_s(y_i|t_i) = P_s(y(t_i))$.

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

The mean welfare of treatment t in state of nature s is denoted by $\mu_{st} \equiv \mathrm{E}_s[u(y(t))]$. The maximum mean welfare achievable in state s is $\max\limits_{t \in T} \mu_{st}$.

After 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 across repeated realizations of the trial is

$$\int_\Psi \sum_{t \in T} \big(\mu_{st} \delta(t, \psi)\big)\, dQ_s(\psi) = \sum_{t \in T} \mu_{st} E_s[\delta(t, \psi)],$$

where $E_s[\delta(t, \psi)] = \int_\Psi \delta(t, \psi) dQ_s(\psi)$ is the expected (across samples) fraction of persons assigned to t.

Application of rule δ in state of nature s yields expected loss (regret)

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

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

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

Hypothesis Testing Rules

First consider 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, 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. The t-statistic for comparing the mean outcome of treatment *t* = 2,…,*L* with that of standard care 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. We use the t-distribution for two-arm trials and the Dunnett’s test critical value for multiple comparisons for multi-arm trials.

The 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,…,*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).

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

The Empirical Success Rule

Let $\bar{u}_t = \frac{1}{n_t}\sum_{i:t_i=t} u(y_i)$ denote the average welfare observed in treatment arm t = 1, 2, …, L.

The empirical success rule prescribes the treatment with the largest observed average patient welfare. If there is a tie, all treatments with the largest observed average patient welfare are prescribed with equal probability.

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

Computing near-optimality for two-arm trials with binary outcomes

When computing the results in Table 2, S is the set of all 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. $\psi = (m_1, m_2)$ 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.

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}.

Computing near-optimality for multi-arm trials with binary outcomes

In Table 4, S is the set of all 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. $\psi = (m_1, \ldots, m_L)$ 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. Given each value of $(p_1, \ldots, p_L)$ we simulate a large number of trial outcomes to approximate Q_s. Our computations of the maximum of (A2) proceed in three steps.

(1) 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 $p_1 = a$, $p_2 = b$, $p_3 = p_4 = p_5 = c$, $b > a$, $b > c$.

(2) We conduct a grid search over these two lower-dimensional parameter spaces using 101 possible parameter values from [0, 0.01, …, 1] for a, b, and c. In this step we approximate Q_s by simulating 1,000,000 trial outcomes.

(3) 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.