ELSEVIER

ScienceDirect

Contents lists available at **sciencedirect.com**
Journal homepage: **www.elsevier.com/locate/jval**

Themed Section: COVID-19

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

Charles F. Manski, PhD, Aleksey Tetenov, PhD

ABSTRACT

Objectives: Researchers studying treatment of coronavirus disease 2019 (COVID-19) have reported findings of randomized trials comparing standard care with care augmented by experimental drugs. Many trials have small sample sizes, so estimates of treatment effects are imprecise. Hence, clinicians may find it difficult to decide when to treat patients with experimental drugs. A conventional practice when comparing standard care and an innovation is to choose the innovation only if the estimated treatment effect is positive and statistically significant. This practice defers to standard care as the status quo. We study treatment choice from the perspective of statistical decision theory, which considers treatment options symmetrically when assessing trial findings.

Methods: We use the concept of near-optimality to evaluate criteria for treatment choice. This concept jointly considers the probability and magnitude of decision errors. An appealing criterion from this perspective is the empirical success rule, which chooses the treatment with the highest observed average patient outcome in the trial.

Results: Considering the design of some COVID-19 trials, we show that the empirical success rule yields treatment choices that are much closer to optimal than those generated by prevailing decision criteria based on hypothesis tests.

Conclusion: Using trial findings to make near-optimal treatment choices rather than perform hypothesis tests should improve clinical decision making.

Keywords: COVID-19, decision criteria, near optimality, randomized trials.

VALUE HEALTH. 2021; 24(5):641–647

Introduction

Researchers studying treatment of coronavirus disease 2019 (COVID-19) have reported findings of randomized trials comparing standard care with care augmented by experimental drugs. Many trials have small sample sizes, so estimates of treatment effects are statistically imprecise. Seeing imprecision, clinicians find it difficult to decide when to treat patients with experimental drugs. Whatever criterion one uses, there is some probability that random variation in trial outcomes will lead to prescribing suboptimal treatments.

A conventional practice when comparing standard care and an innovation is to choose the innovation only if the estimated treatment effect is positive and statistically significant. This practice, which defers to standard care as the status quo, is mandated in regulatory drug-approval processes and is used widely elsewhere. To evaluate decision criteria in nonregulatory settings, we use the concept of *near-optimality,* which jointly considers the probability and magnitude of decision errors. An appealing decision criterion from this perspective is the empirical success rule, which chooses the treatment with the highest observed average patient outcome in the trial.

The contributions of this article are both applied and methodological. We apply to recent COVID-19 trials the general methodology for decision-theoretic study of 2-arm trials developed in Manski[1] and Manski and Tetenov.[2-3] We extend the computational reach of the methodology to enable practical analysis of multi-arm trials. We show that the empirical success rule yields results that are much closer to optimal than those generated by prevailing decision criteria based on hypothesis tests.

Background

A core objective of randomized trials is to inform treatment choice. When comparing standard care with an innovation, the prevailing statistical practice has been to conclude that the innovation is better than standard care only if the estimated average treatment effect comparing the innovation with standard care is statistically significant. Equivalently, a test must reject the null hypothesis that the innovation is no better than standard care. Statistical analysis commonly examines predeclared primary and secondary outcomes of a trial in isolation from one another rather than the joint effect of all outcomes. Articles reporting trials often report subgroup findings only when they are statistically significant.

Figure 1 summarizes a well-cited nonregulatory trial[4] comparing standard care for severe COVID-19 with standard care augmented by prescription of lopinavir/ritonavir. A clinician might reasonably view the estimated reductions in median time to clinical improvement and in mortality to be suggestive evidence that treatment with lopinavir/ritonavir is beneficial relative to standard care alone. Yet the study authors conclude: "no benefit was observed with lopinavir/ritonavir treatment beyond standard care."[4(p1787)] This conclusion was reached because the estimated treatment effects were not statistically significant. Subsequently, COVID-19 treatment guidelines issued by National Institute of Health[5] cited the absence of statistical significance when it characterized the study as having negative findings.

Requiring statistical significance to prescribe a treatment innovation shows deference to standard care, placing the burden of proof on the innovation. One might argue that it is reasonable to place the burden on an innovation when standard care is known to yield good patient outcomes, but the effectiveness of the innovation is uncertain. This argument lacks appeal in the COVID-19 setting. Standard care for COVID-19 developed rapidly to cope with an emergency. The versions of standard care administered early in the pandemic were not shown to yield notably good outcomes.

How might clinicians act with imprecise evidence such as in the Cao et al study? Bayesian statisticians have long criticized the use of hypothesis testing to design trials and to make treatment decisions. The literature on Bayesian statistical inference rejects the frequentist foundations of hypothesis testing, arguing for superiority of the Bayesian practice of using sample data to transform a subjective prior distribution on treatment response into a subjective posterior distribution.[6,7] This done, one chooses a treatment to maximize posterior subjective welfare for a specified welfare function.[8-10] The usefulness of performing a trial is expressed by the expected value of information,[11] defined in Meltzer[12] as "the change in expected utility with the collection of information."

The expected value of information provided by a trial crucially depends on the prior distribution placed on treatment response. The Bayesian perspective is compelling when a decision maker feels able to assert a credible prior distribution. However, Bayesian statisticians have long struggled to provide guidance on specification of priors, and the matter continues to be controversial. See, for example, the spectrum of views expressed by the authors and discussants of Spiegelhalter et al[6] and Manski.[13] The controversy suggests that inability to express a credible prior is common in actual decision settings.

When it is difficult to place a credible subjective distribution on treatment response, a reasonable way to make treatment choices is to use a decision rule that achieves uniformly satisfactory results, whatever the true distribution of treatment response may be. This motivates use of the near-optimality concept to evaluate trial findings.

Measuring the Near-Optimality of Criteria Using Trial Data to Choose a Treatment

The results in any randomized trial have random variation. Whatever criterion one uses to make treatment decisions based on trial results, there is some probability that random variation will lead to prescribing a suboptimal treatment to patients. Considering the probability of error alone is insufficient. The same error probability should be less tolerable when the impact of suboptimal treatment on patient welfare is larger. To evaluate decision criteria, we use the concept of *near-optimality*, which jointly considers the probability of errors and their magnitudes. This concept was proposed abstractly by Savage[14] and has been studied in the context of treatment choice with trial data by Manski,[1] Manski and Tetenov,[2-3] and others.

The concept is as follows. Consider specified possible values for average patient outcomes under each treatment. Presuming the common medical focus on average patient outcomes, the ideal clinical decision would prescribe a treatment that maximizes average outcome. Trial data do not reveal the best treatment with certainty, so one cannot achieve this ideal. Suppose then that one applies some decision criterion to the data. The criterion may be a hypothesis test or another one that we will introduce shortly.

For every treatment that is not best, we compute the frequentist probability that it would be prescribed when the criterion is applied to the results of a trial. We multiply this error probability by the magnitude of the loss from prescribing this treatment, measured by the difference in average patient outcomes compared to the best treatment. This product measures the expected loss from prescribing the inferior treatment, also called its *regret*. The sum of these expected losses across all inferior treatments measures the gap between the ideal of prescribing the best treatment and the reality of having to prescribe the treatment using trial-based estimates subject to random variation.

The aforementioned calculations are made using specified possible values for average patient outcomes with each treatment. However, trial data do not reveal the true values for average patient outcomes; they only enable one to estimate them. The final measurement step is to look across all possible values for average patient outcomes for all treatments to find the values where the expected loss from prescribing inferior treatments is largest. This measures the nearness to optimality of the proposed criterion. Nearness to optimality is also called *maximum regret*. See Appendix A in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.11.019 for a mathematical statement.

Illustrative Application

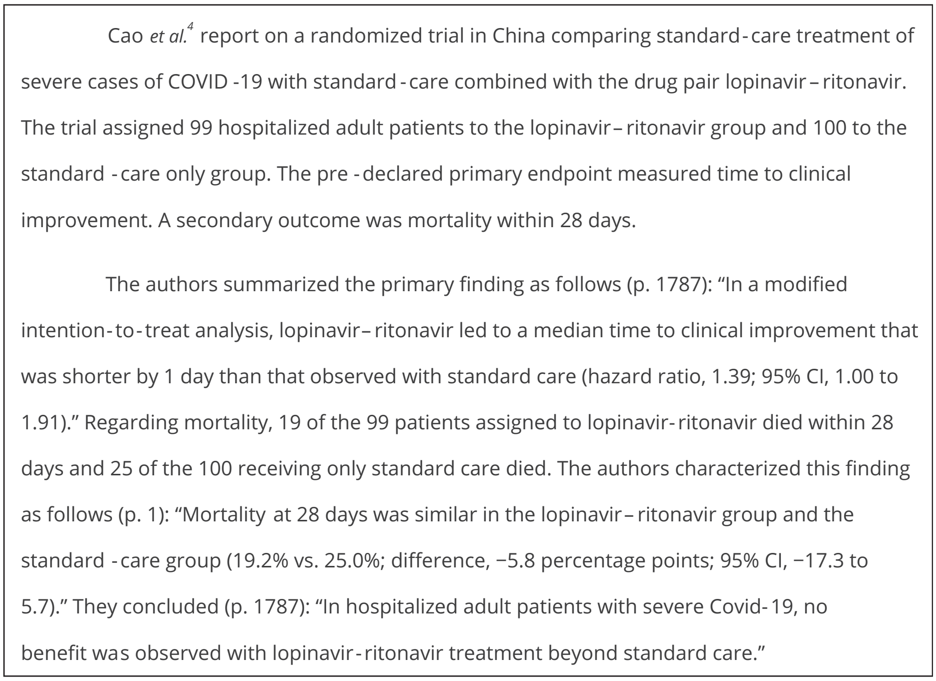
To illustrate measurement of nearness to optimality, Table 1 applies 2 decision criteria to the trial design in Cao et al,[4] which assigned 100 patients to standard care and 99 to care augmented by lopinavir/ritonavir. We focus on 28-day mortality, presumably the most important outcome for patients with severe COVID-19. Each column in the table specifies one scenario for average patient outcomes, combining a mortality rate of standard care, fixed at 0.25, with a mortality rate of the new treatment, ranging from 0.4 to 0.1.

Panel A shows what would happen if the data were used to make treatment decisions with a 2-sided *t* test at 5% level. Thus, the new treatment would be prescribed if the results of the test show the new treatment to be statistically significantly better than standard care. If the new treatment is better, prescribing standard care is an error. The loss from this error is the difference in average patient outcomes.

The table shows that if the new treatment has a mortality rate of 0.15, compared to 0.25 for standard care, a trial with the design of Cao et al[4] will erroneously reach a negative conclusion about the new treatment in 57.4% of trials, leading clinicians to continue using standard care. The magnitude of the error is 0.1, the difference between 0.25 and 0.15. Multiplying the probability of error by its magnitude gives an expected loss of 0.0574.

Suppose instead that the new treatment has mortality rate 0.2. Then the test would reach a negative conclusion about the new

Figure 1. Statistical analysis in a trial comparing treatments for coronavirus disease 2019.



Cao *et al.*[4] report on a randomized trial in China comparing standard-care treatment of severe cases of COVID-19 with standard-care combined with the drug pair lopinavir–ritonavir. The trial assigned 99 hospitalized adult patients to the lopinavir–ritonavir group and 100 to the standard-care only group. The pre-declared primary endpoint measured time to clinical improvement. A secondary outcome was mortality within 28 days.

The authors summarized the primary finding as follows (p. 1787): "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)." Regarding mortality, 19 of the 99 patients assigned to lopinavir-ritonavir died within 28 days and 25 of the 100 receiving only standard care died. The authors characterized this finding as follows (p. 1): "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)." They concluded (p. 1787): "In hospitalized adult patients with severe Covid-19, no benefit was observed with lopinavir-ritonavir treatment beyond standard care."

treatment in 86.8% of trials. While the error probability in this scenario is higher, it is less consequential for clinical outcomes because the difference in mortality rates between treatments is 0.05. Expected loss is 0.868 × 0.05 = 0.0434.

If the new treatment has mortality rate 0.3 (0.05 higher than standard care), the test would reach a positive conclusion about the new treatment only in 0.3% of trials, leading to expected loss of 0.003 × 0.05 = 0.00015. Expected loss is also extremely low in other scenarios where the new treatment has a considerably higher mortality rate than standard care because the probability of type I error of a hypothesis test is dramatically lower than its nominal size.

The nominal size 0.05 of the test is the error probability in the borderline case where the 2 treatments have the same mortality rate. A 2-sided test rejects the null hypothesis if the new treatment performs sufficiently better or worse than standard care. The allowed type I error probability is split between these cases, but rejection of the null hypothesis only leads to prescription of the new treatment in the first case.

We measure nearness to optimality by considering all possible scenarios for the average outcomes of treatments in the trial, which can take any values in the [0, 1] interval, not just the few scenarios illustrated in Table 1. We report nearness to optimality for treatment choice based on *t* tests in 2-arm trials with different sample sizes in Table 2. The table shows that choosing treatments based on a *t* test following a 2-arm trial in which 100 patients receive each treatment (as in Cao et al[4]) achieves near-optimality of 0.071. The maximum value of expected loss across all possible values of average mortality rates occurs when the new treatment has mortality rate 0.548 and standard care has rate 0.661. Then the expected loss (0.661 − 0.548) multiplied by the error probability 0.624 equals 0.071.

Hypothesis tests treat standard care and the new treatment asymmetrically. An appealing alternative decision criterion is the empirical success rule, studied in Manski[1] and Manski and Tetenov.[2,3] This criterion chooses the treatment with the highest observed average patient outcome in the trial, regardless of statistical significance. Whereas hypothesis testing favors standard care and places the burden of proof on innovations, the empirical success rule assesses the evidence on each treatment symmetrically, not distinguishing semantically between standard care and an innovation.

The properties of the empirical success rule are illustrated in panel B of Table 1. If the new treatment has mortality rate 0.2 and standard care has rate 0.25, the empirical success rule will prescribe the new treatment in 78.8% of trials, whereas the testing approach of panel A would only do so in 13.2% of trials. The expected losses when the new treatment is better and when standard care is better are also symmetric.

Table 2 compares near-optimality of the empirical success rule and the test-based decision criterion in 2-arm trials for a wide range of sample sizes. These calculations consider all possible values for the average mortality rates of the 2 treatments. Appendix A in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.11.019 describes the algorithm used to compute near-optimality.

The empirical success rule is about 6 times nearer to optimality than the test-based decision criterion. In a trial with 100 patients in each arm, the empirical success rule achieves near-optimality of 0.012. The maximum value of expected loss occurs when standard care and the new treatment have mortality rates of 0.527 and 0.473. In this case, standard care is erroneously prescribed with probability 0.226. The same expected loss occurs when standard care has mortality rate 0.473 and the new treatment has rate 0.527. Then the new treatment is also erroneously prescribed with probability 0.226.

Good near-optimality properties of the empirical success rule in 2-arm trials are well established in the theoretical literature. Given any specified sample size, the empirical success rule achieves the lowest possible value of near-optimality in trials with

Table 1. Illustrative scenarios for a trial assigning 100 patients to standard care and 99 to a new treatment, as in Cao et al.[4]

Mortality rates:							
Standard care alone	0.25	0.25	0.25	0.25	0.25	0.25	0.25
With new treatment	0.4	0.35	0.3	0.25	0.2	0.15	0.1
Panel A: What happens if treatment decisions are made using a 2-sided 5% hypothesis test							
Trials after which standard care will be prescribed (%)	100.00	99.98	99.70	97.50	86.76	57.36	18.92
Loss from prescribing standard care	0	0	0	0	0.05	0.1	0.15
Trials after which new treatment will be prescribed (%)	0.00	0.02	0.30	2.50	13.24	42.64	81.08
Loss from prescribing new treatment	0.15	0.1	0.05	0	0	0	0
Expected loss:	0.0000	0.0000	0.0002	0.0000	0.0434	0.0574	0.0284
Panel B: What happens if treatment decisions are made using the empirical success rule							
Trials after which standard care will be prescribed (%)	98.95	94.28	79.61	51.64	21.18	4.22	0.26
Loss from prescribing standard care	0	0	0	0	0.05	0.1	0.15
Trials after which new treatment will be prescribed (%)	1.05	5.72	20.39	48.36	78.82	95.78	99.74
Loss from prescribing new treatment	0.15	0.1	0.05	0	0	0	0
Expected loss	0.0016	0.0057	0.0102	0.0000	0.0106	0.0042	0.0004

binary outcomes that assign an equal number of patients to each arm.[15] It does so asymptotically in general trials comparing 2 treatments.[16]

Implications for Clinical Decision Making and Trial Design

Suppose that a clinician were to choose between standard care and standard care augmented with lopinavir/ritonavir based solely on the results of Cao et al,[4] using standard hypothesis testing. As discussed earlier, maximum expected loss relative to optimal treatment is 0.071. Thus, the average mortality rate of these patients could be up to 0.071 higher than under the better of the 2 treatments. (This average is over different possible trial results.) Given the gravity of the patient outcomes at stake, this may be an unacceptably high expected loss in welfare.

There are 2 ways of reducing maximum expected loss: (1) increase sample size and (2) change the way trial results are translated into clinical practice. Table 2 shows that a trial enrolling 4000 patients into each arm, followed by treatment choice using standard hypothesis testing, would achieve near-optimality of 0.0115. About the same level of near-optimality (0.0120) could be achieved by using the empirical success rule in a trial with 100 patients in each arm. Thus, the empirical success rule yields a dramatic improvement in near-optimality relative to testing.

Whether one uses the empirical success rule or a hypothesis test to choose treatments, increasing sample size improves nearness to optimality. Considering 2-arm trials with equal numbers of patients in each arm, Table 2 quantifies the improvement in near-optimality as sample size increases from 20 to 15 000. The literature on testing cautions against designing trials with severely

Table 2. Near-optimality of hypothesis test and empirical success decision rules for 2-arm trials with equal number of patients in each arm.

Sample size per arm	Near-optimality if treatment decisions are made using a 2-sided 5% hypothesis test	Near-optimality if treatment decisions are made using the empirical success rule
20	0.1685	0.0269
30	0.1304	0.0220
50	0.0990	0.0170
100	0.0705	0.0120
200	0.0510	0.0085
500	0.0319	0.0054
1000	0.0228	0.0038
2000	0.0161	0.0027
4000	0.0115	0.0019
5000	0.0102	0.0017
10000	0.0073	0.0012
15000	0.0059	0.0010

small sample sizes because they have low statistical power. We similarly caution that decisions based on the findings of severely small trials may be far from optimal.

Medical research evaluating pharmaceuticals has traditionally shown deference to standard care. Hence, one might question the empirical success rule on the grounds that it evaluates the treatments in the trial symmetrically and thus has the same levels of type I and type II errors. We think symmetric evaluation of standard care and innovations is justified in the COVID-19 setting when considering nonregulatory trials that compare carefully chosen treatments, without a financial conflict of interest, and that report all patient-relevant outcomes. We do not address regulatory trials, whose rules should recognize that the drug approval process may affect the decisions of pharmaceutical firms to perform trials and submit applications for approval of new drugs.[17] For example, loosening statistical criteria for drug approval may induce firms to seek approval for less effective drugs.

It may be reasonable to argue that a risk-averse clinician observing the results of a nonregulatory trial should give standard care the benefit of the doubt if standard care is known to yield good patient outcomes, whereas the effectiveness of an innovation is uncertain. However, this argument is inapplicable when considering COVID-19 treatment early in the pandemic, when the outcomes of standard care were themselves highly uncertain and yet clinicians had to quickly make treatment decisions for severely ill patients. This suggests an ethical symmetry between the possibilities that standard care is better and that standard care augmented by experimental drugs is better. It is logical, then, to evaluate the 2 treatments symmetrically.

Near-Optimality in Multi-arm Trials

Some promising pharmaceutical treatments for COVID-19 have undergone clinical trials. Most are 2-arm trials, comparing an experimental treatment with standard care. It is important for clinicians to learn not only which treatments are better than standard care, but also which new treatments are the most effective.

Running multiple 2-arm trials has a significant drawback when there are concurrently several treatments under investigation: the performance of alternative treatments cannot easily be compared between trials because the populations from which different trials recruit patients usually are not the same. Trials may also differ in the characteristics of the standard care they provide and in the outcomes they report. These problems are addressed by multi-arm trials that randomize the same patients either to standard care or to one of several experimental treatments.

Two large-scale multi-arm trials of treatments for COVID-19 have been initiated. The Recovery Trial[18] in the United Kingdom initially compared standard care with 4 alternatives. The international Solidarity Trial[19] organized by the World Health Organization compares standard care with 5 options. We will consider the initial design of the Recovery Trial, which assigned patients to standard care and alternative treatments in a 2:1:1:1:1 ratio. The Solidarity Trial had balanced assignment of patients to treatments.

The standard way to analyze the results of multi-arm trials has been to compute a t statistic for the difference in average trial outcomes between each new treatment and standard care. Each t statistic is then compared to a critical value adjusted for multiplicity of hypotheses. The aim of this adjustment is to guarantee that in a scenario when all new treatments have the same true average outcome as standard care, there is only a 0.05 probability that any of the differences will be found to be statistically significant in a trial. The Recovery Trial protocol follows this convention and states that Dunnett's test of multiple hypotheses will be used. The intention to use Dunnett's test may have motivated the study team to assign patients in a 2:1:1:1:1 ratio, which has been recommended when applying this test.[20]

Table 3 illustrates how the near-optimality of a decision criterion is evaluated in a multi-arm trial. We consider a trial, similar in design to the Recovery Trial, randomizing 1500 patients: 500 to standard care and all others to 1 of 4 new treatments (250 to each). The table shows what happens in a scenario where the mortality rate of standard care is 0.25 and the mortality rates of treatments A, B, C, and D are 0.15, 0.2, 0.3, and 0.35.

Panel A shows what would happen if the trial data were used to make treatment decisions based on a 2-sided Dunnett's test at the 5% level. We assume that standard care will be prescribed if none of the new treatments has a lower mortality rate that is statistically significantly better. If 1 or more new treatments is statistically significantly better, the new treatment with the lowest mortality rate among them will be prescribed.

Treatment A has the lowest mortality rate in this scenario and will be prescribed after 70.6% of trials. Standard care will be prescribed after 25.7% of trials. Because standard care has a mortality rate that is 0.1 higher than the best treatment (A), this error yields a loss of 0.1. The expected loss from prescribing standard care is the product of the error probability and its magnitude: 0.257 × 0.1 = 0.0257. Treatment B will be prescribed after 3.8% of trials. Because its mortality rate is 0.05 higher than that of the best treatment, the expected loss from prescribing treatment B is 0.038 × 0.05 = 0.0019. Prescribing B does not increase patient mortality rate as much as prescribing standard care, and the expected loss reflects that. Treatments C and D will be prescribed after fewer than 0.01% of trials, and the expected loss from these errors is negligible. Overall expected loss in this scenario is 0.0275, with 0.0257 resulting from prescribing standard care and 0.0019 from prescribing treatment B. Although standard care is only the third-best option, it is prescribed much more frequently than the second-best option (B) due to the status quo deference in hypothesis testing.

Panel B shows what would happen if the empirical success rule were used. Treatment A would be prescribed after 93% of trials. The second-best treatment B would be prescribed after 7% of trials, resulting in expected loss of 0.07 × 0.05 = 0.0035. Standard care would be prescribed only after 0.02% of trials, and treatments C and D after fewer than 0.01% of trials. The overall expected loss when using the empirical success rule in this scenario is 0.0035.

Near-optimality is measured by considering all possible scenarios for the average outcomes of treatments in the trial. Appendix A in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.11.019 describes the algorithm used to compute near-optimality. In Table 4 we compare the near-optimality of prescribing treatments using standard multiple hypothesis testing and of prescribing them using the empirical success rule in 5-arm trials with different sample sizes. We report results both for trials with a 2:1:1:1:1 treatment-assignment ratio (as in the Recovery Trial) and for trials with the same total sample size, but balanced assignment of patients to treatments. In each case considered, the empirical success rule is more than 3 times nearer to optimality than the test-based decision criterion.

Table 4 shows that use of Dunnett's Test with the (500:250:250:250:250) treatment-assignment rates of the Recovery Trial yields near-optimality value 0.0532. The table shows that the empirical success rule with a much smaller sample size (assignment rates 100:50:50:50:50) yields a better near-optimality value of 0.0362.

Table 3. Illustrative scenario for a multi-arm clinical trial assigning 500 patients to receive standard care and 250 patients each to 4 alternative treatments.

	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 2-sided Dunnett's test at 5% significance					
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					
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**

Near-Optimality of the Empirical Success Rule With Patient-Specific Treatment and Multiple Outcomes

The calculations of near-optimality in Tables 1 through 4 concern relatively simple settings where patients are observationally identical and trial outcomes are binary, such as mortality. In clinical practice, trial outcomes may take multiple values. For example, trials of COVID-19 drugs may report mortality outcomes and time to recovery for patients who survive. Patients who vary in age, gender, and comorbidities may vary in their response to treatment.

It has been common in analysis of trial data to designate primary and secondary outcomes. The latter are often called side effects. Research articles focus attention on the primary outcome. This 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 secondary effects of treatments vary markedly across treatments, it is more reasonable to consider how the primary and secondary outcomes jointly determine patient welfare.

This is easy to accomplish with the empirical success rule. Methodological research has shown how to compute or bound the near-optimality of the empirical success rule in a broad range of settings. Appendix B in Supplemental Materials found at https://doi.org/10.1016/j.jval.2020.11.019 summarizes the findings.

Discussion

A central objective of clinical trials is to inform treatment choice. Yet researchers analyzing trial data have used concepts of statistical inference whose foundations are distant from treatment choice. It has been common to use hypothesis tests to choose treatments. In earlier work, we have proposed evaluation of decision criteria by near-optimality. Here we apply the concept to analyze findings of trials comparing COVID-19 treatments. We find that the empirical success rule performs much better than hypothesis testing.

Of course, use of the empirical success rule does not guarantee that the optimal treatment is always chosen. No decision criterion can achieve this ideal with finite trial data. Evaluation of criteria by near-optimality appropriately recognizes how the probability and magnitude of errors in decision making combine to affect patient welfare. Increasing sample size decreases error probabilities and, hence, increases nearness to optimality.

For simplicity, we have considered trials having full internal and external validity. Internal validity may be compromised by

Table 4. Near-optimality of multiple hypothesis testing and empirical success decision rules for 5-arm trials with specified sample sizes.

Sample sizes for each arm	Near-optimality if treatment decisions are made using a 2-sided 5% Dunnett's test	Near-optimality if treatment decisions are made using the empirical success rule
100:50:50:50:50	0.1224	0.0362
60:60:60:60:60	0.1251	0.0343
200:100:100:100:100	0.0855	0.0256
120:120:120:120:120	0.0859	0.0243
500:250:250:250:250	0.0532	0.0160
300:300:300:300:300	0.0563	0.0153
1000:500:500:500:500	0.0380	0.0112
600:600:600:600:600	0.0390	0.0107
2000:1000:1000:1000:1000	0.0274	0.0080
1200:1200:1200:1200:1200	0.0291	0.0076

noncompliance and loss to follow-up. External validity may be compromised by measurement of surrogate outcomes and by administration of treatments to types 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 numerical calculations made in this article require modification.

A limitation of this article is that it only considers treatment choice using data from 1 trial. In practice, a clinician may learn the findings of multiple trials and may also be informed by observational data. The concept of near-optimality is well defined in these more complex settings, but methods for practical application are yet to be developed.

A further issue beyond the scope of this article concerns the dynamics of treatment choice when new trial data and observational evidence may emerge in the future. Dynamics is also a consideration in the design of trials, whose rules may include provision for early stopping as results emerge. The concept of near-optimality is extendable to dynamic settings. However, methodology for application is yet to be developed.

Dynamic analysis of treatment choice made with hypothesis tests may be especially difficult to perform, because testing views standard care and new treatments asymmetrically. As new evidence accumulates over time, the consensus designation of standard care may change, leading to a change in the null hypothesis when new trials are evaluated. The implications for patient welfare are unclear.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2020.11.019.

Article and Author Information

Accepted for Publication: November 17, 2020

Published Online: March 25, 2021

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

Author Affiliations: Department of Economics and Institute for Policy Research, Northwestern University, Evanston, IL, USA (Manski); Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland (Tetenov).

Correspondence: Charles F. Manski, Department of Economics, Northwestern University, 2211 Campus Drive, Evanston, IL 60208-2600, USA. Email: cfmanski@northwestern.edu

Author Contributions: *Concept and design*: Manski, Tetenov
Acquisition of data: Manski
Analysis and interpretation of data: Manski
Drafting of the manuscript: Manski, Tetenov
Critical revision of the paper for important intellectual content: Manski
Statistical analysis: Manski, Tetenov
Obtaining funding: Tetenov

Conflict of Interest Disclosures: The authors reported no conflicts of interest.

Funding/Support: This work was supported by grant 100018-192580 from the Swiss National Science Foundation.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Acknowledgment: We have benefited from the comments of Michael Gmeiner, Valentyn Litvin, Francesca Molinari, John Mullahy, and anonymous reviewers.

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.
4. Cao B, Wang Y, Wen D, et al. A trial of lopinavir–ritonavir in adults hospitalized with severe COVID-19. *N Engl J Med*. 2020;382:1787–1799.
5. National Institute of Health. Potential antiviral drugs under evaluation for the treatment of COVID-19. https://www.covid19treatmentguidelines.nih.gov/antiviral-therapy/; 2020. Accessed September 27, 2020.
6. Spiegelhalter D, Freedman L, Parmar M. Bayesian approaches to randomized trials. (with discussion). *J R Stat Soc Series A*. 1994;157(3):357–416.
7. Spiegelhalter D. Incorporating Bayesian ideas into health-care evaluation. *Stat Sci*. 2004;19(1):156–174.
8. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ*. 1998;18(3):341–364.
9. Lindley D. Decision analysis and bioequivalence trials. *Stat Sci*. 1998;13(2):136–141.
10. Berry D. Bayesian statistics and the efficiency and ethics of clinical trials. *Stat Sci*. 2004;19(1):175–187.
11. Claxton K, Posnett J. An economic approach to clinical trial design and research priority-setting. *J Health Econ*. 1996;5(6):513–524.
12. Meltzer D. Addressing uncertainty in medical cost-effectiveness: analysis implications of expected utility maximization for methods to perform sensitivity analysis and the use of cost-effectiveness analysis to set priorities for medical research. *J Health Econ*. 2001;20(1):109–129.
13. Manski C. Reasonable patient care under uncertainty. Health Econ. 27(10): 1397-1421.
14. Savage L. The theory of statistical decision. *J Amer Stat Assoc*. 1951;46(253):55–67.
15. Stoye J. Minimax regret treatment choice with finite samples. *J Econom*. 2009;151(1):70–81.
16. Hirano K, Porter J. Asymptotics for statistical treatment rules. *Econometrica*. 2009;77(5):1683–1701.
17. Tetenov A. An Economic Theory of Statistical Testing, 2016. Cemmap working paper CWP50/16.
18. Nuffield Department of Population Health. Recovery. https://www.recoverytrial.net/. Accessed November 6, 2020.
19. World Health Organization. “Solidarity” clinical trial for COVID-19 treatments. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov/solidarity-clinical-trial-for-covid-19-treatments. Accessed November 6, 2020.
20. Dunnett C. New tables for multiple comparisons with a control. *Biometrics*. 1964;20(3):482–491.