Sufficient trial size to inform clinical practice

Charles F. Manski[a,1] and Aleksey Tetenov[b,c]

[a]Department of Economics and Institute for Policy Research, Northwestern University, Evanston, IL 60208; [b]Department of Economics, University of Bristol, Bristol BS8 1TU, United Kingdom; and [c]Collegio Carlo Alberto, Moncalieri (TO) 10024, Italy

Contributed by Charles F. Manski, July 23, 2016 (sent for review May 20, 2016; reviewed by Keisuke Hirano and David Meltzer)

Medical research has evolved conventions for choosing sample size in randomized clinical trials that rest on the theory of hypothesis testing. Bayesian statisticians have argued that trials should be designed to maximize subjective expected utility in settings of clinical interest. This perspective is compelling given a credible prior distribution on treatment response, but there is rarely consensus on what the subjective prior beliefs should be. We use Wald's frequentist statistical decision theory to study design of trials under ambiguity. We show that ε-optimal rules exist when trials have large enough sample size. An ε-optimal rule has expected welfare within ε of the welfare of the best treatment in every state of nature. Equivalently, it has maximum regret no larger than ε. We consider trials that draw predetermined numbers of subjects at random within groups stratified by covariates and treatments. We report exact results for the special case of two treatments and binary outcomes. We give simple sufficient conditions on sample sizes that ensure existence of ε-optimal treatment rules when there are multiple treatments and outcomes are bounded. These conditions are obtained by application of Hoeffding large deviations inequalities to evaluate the performance of empirical success rules.

clinical trials | sample size | medical decision making | near optimality

A core objective of randomized clinical trials (RCTs) comparing alternative medical treatments is to inform treatment choice in clinical practice. However, the conventional practice in designing trials has been to choose a sample size that yields specified statistical power. Power, a concept in the statistical theory of hypothesis testing, is at most loosely connected to effective treatment choice.

This paper develops an alternative principle for trial design that aims to directly benefit medical decision making. We propose choosing a sample size that enables implementation of near-optimal treatment rules. Near optimality means that treatment choices are suitably close to the best that could be achieved if clinicians were to know with certainty mean treatment response in their patient populations. We report exact results for the case of two treatments and binary outcomes. We derive simple formulas to compute sufficient sample sizes in clinical trials with multiple treatments.

Whereas our immediate concern is to improve the design of RCTs, our work contributes more broadly by adding to the reasons why scientists and the general public should question the hegemony of hypothesis testing as a methodology used to collect and analyze sample data. It has become common for scientists to express concern that evaluation of empirical research by the outcome of statistical hypothesis tests generates publication bias and diminishes the reproducibility of findings. See, for example, ref. 1 and the recent statement by the American Statistical Association (2). We call attention to a further deficiency of testing. In addition to providing an unsatisfactory basis for evaluation of research that uses sample data, testing also is deficient as a basis for the design of data collection.

Background

The Conventional Practice. The conventional use of statistical power calculations to set sample size in RCTs derives from the presumption that data on outcomes in a classical trial with perfect validity will be used to test a specified null hypothesis against an alternative. A common practice is to use the outcome of a hypothesis test to recommend whether a patient population should receive a status quo treatment or an innovation. The usual null hypothesis is that the innovation is no better than the status quo and the alternative is that the innovation is better. If the null hypothesis is not rejected, it is recommended that the status quo treatment should continue to be used. If the null is rejected, it is recommended that the innovation should replace the status quo as the treatment of choice.

The standard practice has been to perform a test that fixes the probability of rejecting the null hypothesis when it is correct, called the probability of a type I error. Then sample size determines the probability of rejecting the alternative hypothesis when it is correct, called the probability of a type II error. The power of a test is defined as one minus the probability of a type II error. The convention has been to choose a sample size that yields specified power at some value of the effect size deemed clinically important.

The US Food and Drug Administration (FDA) uses such a test to approve new treatments. A pharmaceutical firm wanting approval of a new drug (the innovation) performs RCTs that compare the new drug with an approved drug or placebo (the status quo). An FDA document providing guidance for the design of RCTs evaluating new medical devices states that the probability of a type I error is conventionally set to 0.05 and that the probability of a type II error depends on the claim for the device but should not exceed 0.20 (3). The International Conference on Harmonisation has provided similar guidance for the design of RCTs evaluating pharmaceuticals, stating (ref. 4, 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

Significance

A core objective of trials comparing alternative medical treatments is to inform treatment choice in clinical practice, and yet conventional practice in designing trials has been to choose a sample size that yields specified statistical power. Power, a concept in the theory of hypothesis testing, is at most loosely connected to effective treatment choice. This paper develops an alternative principle for trial design that aims to directly benefit medical decision making. We propose choosing a sample size that enables implementation of near-optimal treatment rules. Near optimality means that treatment choices are suitably close to the best that could be achieved if clinicians were to know with certainty mean treatment response in their patient populations.

Author contributions: C.F.M. and A.T. designed research, performed research, analyzed data, and wrote the paper.

Reviewers: K.H., University of Arizona; and D.M., University of Chicago.

The authors declare no conflict of interest.

This work was presented in part on September 15, 2015 at the Econometrics Seminar of the Department of Economics, Cornell University, Ithaca, NY.

[1]To whom correspondence should be addressed. Email: cfmanski@northwestern.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1612174113/-/DCSupplemental.

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 conventional error probabilities are called "underpowered" and are regularly criticized as scientifically useless and medically unethical. For example, Halpern et al. (ref. 5, p. 358) write "Because such studies may not adequately test the underlying hypotheses, they have been considered 'scientifically useless' and therefore unethical in their exposure of participants to the risks and burdens of human research." Ones with samples larger than needed to achieve conventional error probabilities are called "overpowered" and are sometimes criticized as unethical. For example, Altman (ref. 6, p. 1336) writes "A study with an overlarge sample may be deemed unethical through the unnecessary involvement of extra subjects and the correspondingly increased costs."

Deficiencies of Using Statistical Power to Choose Sample Size. There are multiple reasons why choosing sample size to achieve specified statistical power may yield unsatisfactory results for medical decisions. These include the following:

i) Use of conventional asymmetric error probabilities: As discussed above, it has been standard to fix the probability of type I error at 5% and the probability of type II error for a clinically important alternative at 10–20%, which implies that the probability of type II error reaches 95% for alternatives close to the null. The theory of hypothesis testing gives no rationale for selection of these conventional error probabilities. In particular, 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.

ii) Inattention to magnitudes of losses when errors occur: A clinician should care about more than the probabilities of types 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 welfare difference into account.

iii) Limitation to settings with two treatments: A clinician often chooses among several treatments and many clinical trials compare more than two treatments. However, the standard theory of hypothesis testing contemplates only choice between two treatments. Statisticians have struggled to extend it to deal sensibly with comparisons of multiple treatments (7, 8).

Bayesian Trial Design and Treatment Choice. With these deficiencies in mind, Bayesian statisticians have long criticized the use of hypothesis testing to design trials and 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. See, for example, refs. 9 and 10.

The literature on Bayesian statistical decision theory additionally argues that the purpose of trials is to improve medical decision making and concludes that trials should be designed to maximize subjective expected utility in decision problems of clinical interest. The usefulness of performing a trial is expressed by the expected value of information (11), defined succinctly in Meltzer (ref. 12, p. 119) as "the change in expected utility with the collection of information." The Bayesian value of information provided by a trial crucially depends on the subjective prior distribution. The sample sizes selected in Bayesian trials may differ from those motivated by testing theory. See, for example, refs. 13 and 14.

The Bayesian perspective is compelling when a decision maker feels able to place a credible 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. See, for example, the spectrum of views expressed by the authors and discussants of ref. 9. The controversy suggests that inability to express a credible prior is common in actual decision settings.

Uniformly Satisfactory Trial Design and Treatment Choice with the Minimax-Regret Criterion. 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. There are multiple ways to formalize the idea of uniformly satisfactory results. One prominent idea motivates the minimax-regret (MMR) criterion.

Minimax regret was first suggested as a general principle for decision making under uncertainty by Savage (15) within an essay commenting on the seminal Wald (16) development of statistical decision theory. Wald considered the broad problem of using sample data to make decisions when one has incomplete knowledge of the choice environment, called the state of nature. He recommended evaluation of decision rules as procedures, specifying how a decision maker would use whatever data may be realized. In particular, he proposed measurement of the mean performance of decision rules across repetitions of the sampling process. This method grounds the Wald theory in frequentist rather than Bayesian statistical thinking. See refs. 17 and 18 for comprehensive expositions.

Considering the Wald framework, Savage defined the regret associated with choice of a decision rule in a particular state of nature to be the mean loss in welfare that would occur across repeated samples if one were to choose this rule rather than the one that is best in this state of nature. The actual decision problem requires choice of a decision rule without knowing the true state of nature. The decision maker can evaluate a rule by the maximum regret that it may yield across all possible states of nature. He can then choose a rule that minimizes the value of maximum regret. Doing so yields a rule that is uniformly satisfactory in the sense of yielding the best possible upper bound on regret, whatever the true state of nature may be.

It is important to understand that maximum regret as defined by Savage is computed *ex ante*, before one chooses an action. It should not be confused with the familiar psychological notion of regret, which a person may perceive *ex post* after choosing an action and observing the true state of nature.

A decision made by the MMR criterion is invariant with respect to increasing affine transformations of welfare, but it may vary when welfare is transformed nonlinearly. The MMR criterion shares this property with expected utility maximization.

The MMR criterion is sometimes confused with the maximin criterion. A decision maker using the maximin criterion chooses an action that maximizes the minimum welfare that might possibly occur. Someone using the MMR criterion chooses an action that minimizes the maximum loss to welfare that can possibly result from not knowing the welfare function. Whereas the maximin criterion considers only the worst outcome that an action may yield, MMR considers the worst outcome relative to what is achievable in a given state of nature. Savage (15), when introducing the MMR criterion, distinguished it sharply from maximin, writing that the latter criterion is "ultrapessimistic" whereas the former is not.

Since the early 2000s, various authors have used the MMR criterion to study how a decision maker might use RCT data to subsequently choose treatments for the members of a population (19–27). In these studies, the decision maker's objective has been expressed as maximization of a welfare function that sums treatment outcomes across the population. For example, the objective may be to maximize the 5-y survival rate of a population

of patients with cancer or the average number of quality-adjusted life years of a population with a chronic disease.

The MMR criterion is applicable in general settings with multiple treatments. Regret is easiest to explain when there are two treatments, say A and B. If treatment A is better, regret is the probability of a type I error (choosing B) times the magnitude of the resulting loss in population welfare due to assigning the inferior treatment. Symmetrically, if treatment B is better, regret is the probability of a type II error (choosing A) times the magnitude of the resulting loss in population welfare due to foregoing the superior treatment. In contrast to the use of hypothesis testing to choose a treatment, the MMR criterion views types I and II error probabilities symmetrically and it assesses the magnitudes of the losses that errors produce.

Whereas the work cited above has used the MMR criterion to guide treatment choice after a trial has been performed, the present paper uses it to guide the design of RCTs. We focus on classical trials possessing perfect validity that compare alternative treatments relevant to clinical practice. Treatments may include placebo if it is a relevant clinical option or if it is considered equivalent to prescribing no treatment (28, 29). In particular, we study trials that draw subjects at random within groups of predetermined size stratified by covariates and treatments. *Trials Enabling Near-Optimal Treatment Rules* summarizes the major findings. *Supporting Information* provides underlying technical analysis.

Trials Enabling Near-Optimal Treatment Rules

General Ideas. An ideal objective for trial design would be to collect data that enable subsequent implementation of an optimal treatment rule in a population of interest—one that always selects the best treatment, with no chance of error. Optimality is too strong a property to be achievable with trials having finite sample size, but near-optimal rules exist when classical trials with perfect validity have large enough size.

Given a specified $\varepsilon > 0$, an ε-optimal rule is one whose mean performance across samples is within ε of the welfare of the best treatment, whatever the true state of nature may be. Equivalently, an ε-optimal rule has maximum regret no larger than ε. Thus, an ε-optimal rule exists if and only if the MMR rule has maximum regret no larger than ε.

Choosing sample size to enable existence of ε-optimal treatment rules provides an appealing criterion for design of trials that aim to inform treatment choice. Implementation of the idea requires specification of a value for ε. The need to choose an effect size of interest when designing trials already arises in conventional practice, where the trial planner must specify the alternative hypothesis to be compared with the null. A possible way to specify ε is to make it equal the minimum clinically important difference (MCID) in the average treatment effect comparing alternative treatments.

Medical research has long distinguished between the statistical and the clinical significance of treatment effects (30). Although the idea of clinical significance has been interpreted in various ways, many writers call an average treatment effect clinically significant if its magnitude is greater than a specified value deemed minimally consequential in clinical practice. The International Conference on Harmonisation (ICH) put it this way (ref. 4, p. 1923): "The treatment difference to be detected may be based on a judgment concerning the minimal effect which has clinical relevance in the management of patients."

Research articles reporting trial findings sometimes pose particular values of MCIDs when comparing alternative treatments for specific diseases. For example, in a study comparing drug treatments for hypertension, Materson et al. (31) defined the outcome of interest to be the fraction of subjects who achieve a specified threshold for blood pressure. The authors took the MCID to be the fraction 0.15, stating that this fraction is "the difference specified in the study design to be clinically important," and reported groups of drugs "whose effects do not differ from each other by more than 15 percent" (ref. 31, p. 916).

Findings with Binary Outcomes, Two Treatments, and Balanced Designs. Determination of sample sizes that enable near-optimal treatment is simple in settings with binary outcomes (coded 0 and 1 for simplicity), two treatments, and a balanced design that assigns the same number of subjects to each treatment group. Table 1 provides exact computations of the minimum sample size that enables ε optimality when a clinician uses one of three different treatment rules, for various values of ε.

The first column in Table 1 shows the minimum sample size (per treatment arm) that yields ε optimality when a clinician uses the empirical success (ES) rule to make a treatment decision. The ES rule chooses the treatment with the better average outcome in the trial. The rule assigns half the population to each treatment if there is a tie. It is known that the ES rule minimizes the maximum regret rule in settings with binary outcomes, two treatments, and balanced designs (25).

The second and third columns in Table 1 display the minimum sample sizes that yield ε optimality of rules based on one-sided 5% and 1% hypothesis tests. There is no consensus on what hypothesis test should be used to compare two proportions. We report results based on the widely used one-sided two-sample z test, which is based on an asymptotic normal approximation (32).

The findings are remarkable. A sample as small as 2 observations per treatment arm makes the ES rule ε optimal when $\varepsilon = 0.1$ and a sample of size 145 suffices when $\varepsilon = 0.01$. The minimum sample sizes required for ε optimality of the test rules are orders of magnitude larger. If the z test of size 0.05 is used, a sample of size 33 is required when $\varepsilon = 0.1$ and 3,488 when $\varepsilon = 0.01$. The sample sizes have to be more than double these values if the z test of size 0.01 is used.

Fig. 1 illustrates the difference between error probabilities and regret incurred by the ES rule and the 5% z-test rule for a sample size of 145 per arm, the minimum sample size yielding ε optimality when $\varepsilon = 0.01$. Fig. 1, *Upper* shows how the probability of error varies with the effect size for all possible distributions of treatment response with effect sizes in the range [−0.5, 0.5]. Fig. 1, *Lower* displays the regret (probability of error times the effect size) of the same treatment rules. Maximum regret occurs at intermediate effect sizes. For small effect sizes, regret is small because choosing the wrong treatment is not clinically significant. Regret is also small for large effect sizes, because the probability of error eventually starts declining rapidly with the effect size. Traditional power calculations are not informative about the maximum regret of a test-based rule. Two red vertical lines in Fig. 1 mark effect sizes at which the z test has at least 80% and 90% power. Neither size corresponds to the effect size where regret is maximal.

Findings with Bounded Outcomes and Multiple Treatments. In principle, the existence of ε-optimal treatment rules under any design can be determined by computing the maximum regret of the minimax-regret rule. In practice, determination of the minimax-regret rule and its maximum regret may be burdensome. To date,

Table 1. Minimum sample sizes per treatment enabling ε-optimal treatment choice: binary outcomes, two treatments, balanced designs

ε	ES rule	One-sided 5% z test	One-sided 1% z test
0.01	145	3,488	7,963
0.03	17	382	879
0.05	6	138	310
0.10	2	33	79
0.15	1	16	35

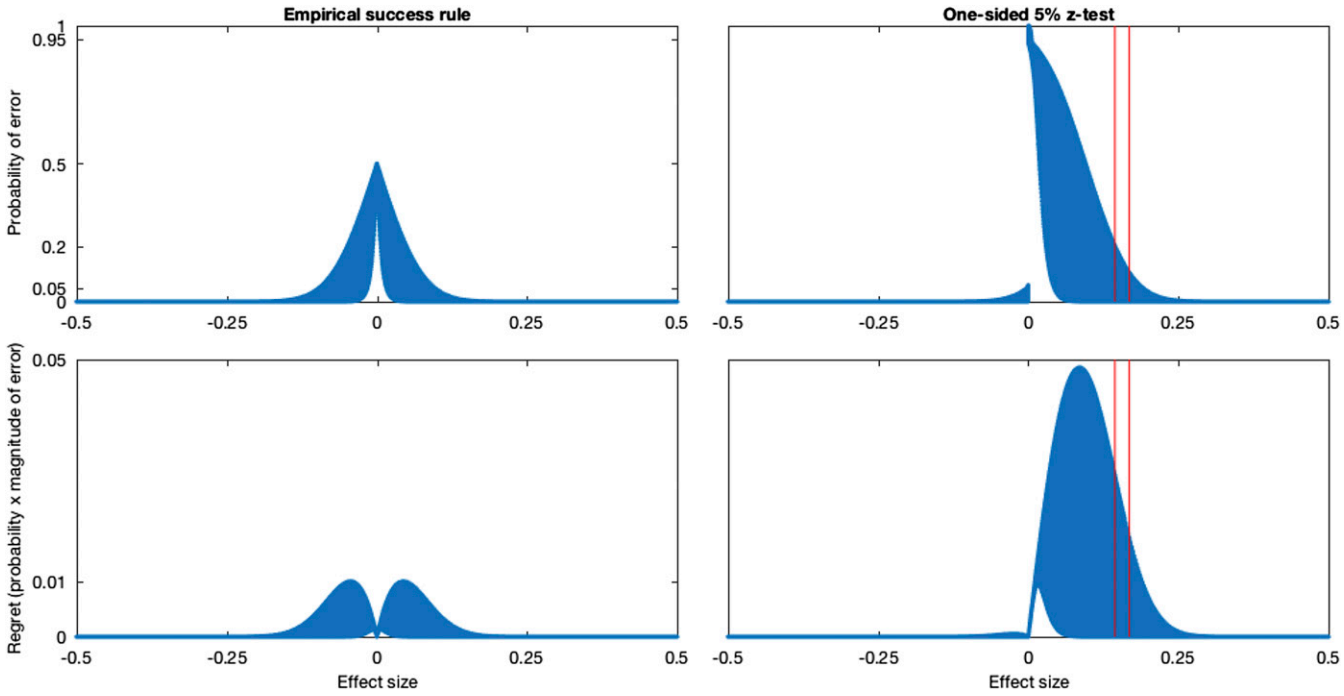

Fig. 1. Error probabilities and regret for empirical success and one-sided 5% *z*-test rules.

exact minimax-regret decision rules have been derived only for the case of two treatments with equal or nearly equal sample sizes (24–26). Hence, it is useful to have simple sufficient conditions that ensure existence of ε-optimal rules more generally. The conditions we derive below hold in all settings where outcomes are bounded. Our findings apply to situations in which there are multiple treatments, not just two. They also apply when trials stratify patients into groups with different observable covariates, such as demographic attributes and risk factors.

To show that a specified trial design enables ε-optimal treatment rules, it suffices to consider a particular rule and to show that this rule is ε optimal when used with this design. We focus on empirical success rules for both practical and analytical reasons. Choosing a treatment with the highest reported mean outcome is a simple and plausible way in which a clinician may use the results of an RCT. Two analytical reasons further motivate interest in ES rules when outcomes are bounded. First, these rules either exactly or approximately minimize maximum regret in various settings with two treatments when sample size is moderate (25, 26) and asymptotic (23). Second, large-deviations inequalities derived in ref. 33 allow us to obtain informative and easily computable upper bounds on the maximum regret of ES rules applied with any number of treatments. These upper bounds on maximum regret immediately yield sample sizes that ensure an ES rule is ε optimal.

Propositions 1 and *2* (*Supporting Information*) present two alternative upper bounds on the maximum regret of an ES rule. *Proposition 1* extends findings of Manski (19) from two to multiple treatments whereas *Proposition 2* derives a new large-deviations bound for multiple treatments. When the design is balanced, these bounds are

$$(2e)^{-(1/2)}M(K-1)n^{-(1/2)}, \quad [1]$$

$$M(\ln K)^{1/2}n^{-(1/2)}, \quad [2]$$

where n is the sample size per arm, K is the number of treatment arms, and M is the width of the range of possible outcomes. *Proposition 3* (*Supporting Information*) shows that the bounds on maximum regret derived in *Propositions 1* and *2* are minimized by balanced designs. Table S1 gives numerical calculations for $K \leq 7$. *Trials Stratified by Observed Covariates* extends these findings to settings where patients have observable covariates.

Propositions 1 and *2* imply sufficient conditions on sample sizes for ε optimality of ES rules. *Proposition 1* implies that an ES rule is ε optimal if the sample size per treatment arm is at least

$$n \geq (2e)^{-1}(K-1)^2(M/\varepsilon)^2. \quad [3]$$

Proposition 2 implies that an ES rule is ε optimal if the sample size per treatment arm is at least

$$n \geq \ln K(M/\varepsilon)^2. \quad [4]$$

We find that when the design is balanced, *Proposition 1* provides a tighter bound than *Proposition 2* for two or three treatments. *Proposition 2* gives a tighter bound for four or more treatments.

To illustrate the findings, consider the Materson et al. (31) study of treatment for hypertension. The outcome is binary with the range of possible outcomes $M = 1$. The study compared seven drug treatments and specified 0.15 as the MCID. We cannot know how the authors of the study, who reported results of traditional hypothesis tests, would have specified ε had they sought to achieve ε optimality. If they were to set $\varepsilon = 0.15$, application of bound **4** shows that an ES rule is ε optimal if the number of subjects per treatment arm is at least $(\ln 7) \cdot (0.15)^{-2} = 86.5$. The actual study has an approximately balanced design, with between 178 and 188 subjects in each treatment arm. Application of bound **2** shows that a study with at least 178 subjects per arm is ε optimal for $\varepsilon = (\ln 7)^{1/2}(178)^{-1/2} = 0.105$.

It is important to bear in mind that *Propositions 1* and *2* imply only simple sufficient conditions on sample sizes for ε optimality of ES rules, not necessary ones. These sufficient conditions use only the weak assumption that outcomes are bounded and they rely on Hoeffding large-deviations inequalities for bounded outcomes. In the special case with binary outcomes and two treatments and a balanced design, the sufficient sample sizes

provided by *Proposition 1* are roughly 10 times the size of the exact minimum sample sizes, depending on the value of ε. This result strongly suggests that it is worthwhile to compute exact minimum sample sizes whenever it is tractable to do so.

Trials Stratified by Observed Covariates. Clinical trials often stratify participants by observable covariates, such as demographic attributes and risk factors, and report trial results separately for each group. We consider ε optimality of the ES rule that assigns individuals with covariates ξ to the treatment that yielded the highest average outcome among trial participants with covariates ξ.

There are at least two reasonable ways that a planner may wish to evaluate ε optimality in this setting. First, he may want to achieve ε optimality within each covariate group. This interpretation requires no new analysis. The planner should simply define each covariate group to be a separate population of interest and then apply the analysis of *Findings with Binary Outcomes, Two Treatments, and Balanced Designs* and *Findings with Bounded Outcomes and Multiple Treatments* to each group. The design that achieves group-specific ε optimality with minimum total sample size equalizes sample sizes across groups.

Alternatively, the planner may want to achieve ε optimality within the overall population, without requiring that it be achieved within each covariate group. Bounds **1** and **2** extend to the setting with covariates. With a balanced design assigning n_ξ individuals from covariate group ξ to each treatment, the maximum regret of an ES rule is bounded above by

$$(2e)^{-(1/2)}M(K-1)\sum_{\xi\in X}P(x=\xi)(n_\xi)^{-(1/2)}, \quad [5]$$

$$M(\ln K)^{1/2}\sum_{\xi\in X}P(x=\xi)(n_\xi)^{-(1/2)}. \quad [6]$$

The design that minimizes bound **5** or **6** for a given total sample size generally neither equalizes sample sizes across groups nor makes them proportional to the covariate distribution $P(x=\xi)$. Instead, the relative sample sizes for any pair (ξ, ξ') of covariate values have the approximate ratio

$$n_\xi/n_{\xi'} = [P(x=\xi)/P(x=\xi')]^{2/3}. \quad [7]$$

Such trial designs make the covariate-specific sample size increase with the prevalence of the covariate group in the population, but less than proportionately. Covariate-specific maximum regret commensurately decreases with the prevalence of the covariate group.

Conclusion

Choosing sample sizes in clinical trials to enable near-optimal treatment rules would align trial design directly with the objective of informing treatment choice. In contrast, the conventional practice of choosing sample size to achieve specified statistical power in hypothesis testing is only loosely related to treatment choice. Our work adds to the growing concern of scientists that hypothesis testing provides an unsuitable methodology for collection and analysis of sample data.

We share with Bayesian statisticians who have written on trial design the objective of informing treatment choice. We differ in our application of the frequentist statistical decision theory developed by Wald, which does not require that one place a subjective probability distribution on treatment response. We use the concept of ε optimality, which is equivalent to having maximum regret no larger than ε.

There are numerous potentially fruitful directions for further research of the type initiated here. One is analysis of other types of trials. We have focused on trials that draw subjects at random within groups of predetermined size stratified by covariates and treatments. With further work, the ideas developed here should be applicable to trials where the numbers of subjects who have particular covariates and receive specific treatment are *ex ante* random rather than predetermined.

Our analysis assumed no prior knowledge restricting the variation of response across treatments and covariates. This assumption, which has been traditional in frequentist study of clinical trials, is advantageous in the sense that it yields generally applicable findings. Nevertheless, it is unduly conservative in circumstances where some credible knowledge of treatment response is available. One may, for example, think it credible to maintain some assumptions on the degree to which treatment response may vary across treatments or covariate groups. When such assumptions are warranted, it may be valuable to impose them.

We mentioned at the outset that medical conventions for choosing sample size pertain to classical trials possessing perfect validity. However, practical trials usually have only partial validity. For example, the experimental sample may be representative only of a part of the target treatment population, because experimental subjects typically are persons who meet specified criteria and who consent to participate in the trial. Due to this and other reasons, experimental data may only partially identify treatment response in the target treatment population. The concept of ε optimality extends to such situations.

Finally, we remark that our analysis followed the long-standing practice in medical research of evaluating trial designs by their informativeness about treatment response, without consideration of the cost of conducting trials. The concept of ε optimality can be extended to recognize trial cost as a determinant of welfare.

ACKNOWLEDGMENTS. We have benefited from the comments of Joerg Stoye and from the opportunity to present this work in a seminar at Cornell University.

1. Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* 2(8):e124.
2. Wasserstein R, Lazar N (2016) The ASA's statement on p-values: Context, process, and purpose. *Am Stat* 70(2):129–133.
3. US Food and Drug Administration (1996) *Statistical Guidance for Clinical Trials of Nondiagnostic Medical Devices*. Available at www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm106757.htm. Accessed August 24, 2016.
4. International Conference on Harmonisation (1999) ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonized tripartite guideline. *Stat Med* 18(15):1905–1942.
5. Halpern SD, Karlawish JH, Berlin JA (2002) The continuing unethical conduct of underpowered clinical trials. *JAMA* 288(3):358–362.
6. Altman DG (1980) Statistics and ethics in medical research: III. How large a sample? *BMJ* 281(6251):1336–1338.
7. Dunnett C (1955) A multiple comparison procedure for comparing several treatments with a control. *JASA* 50(272):1096–1121.
8. Cook R, Farewell V (1996) Multiplicity considerations in the design and analysis of clinical trials. *J R Stat Soc Ser A* 159(1):93–110.
9. Spiegelhalter D, Freedman L, Parmar M (1994) Bayesian approaches to randomized trials (with discussion). *J R Stat Soc Ser A* 157(3):357–416.
10. Spiegelhalter D (2004) Incorporating Bayesian ideas into health-care evaluation. *Stat Sci* 19(1):156–174.
11. Claxton K, Posnett J (1996) An economic approach to clinical trial design and research priority-setting. *Health Econ* 5(6):513–524.
12. Meltzer D (2001) 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* 20(1):109–129.
13. Cheng Y, Su F, Berry D (2003) Choosing sample size for a clinical trial using decision analysis. *Biometrika* 90(4):923–936.
14. Berry D (2004) Bayesian statistics and the efficiency and ethics of clinical trials. *Stat Sci* 19(1):175–187.
15. Savage L (1951) The theory of statistical decision. *JASA* 46(253):55–67.
16. Wald A (1950) *Statistical Decision Functions* (Wiley, New York).
17. Ferguson T (1967) *Mathematical Statistics: A Decision Theoretic Approach* (Academic, New York).
18. Berger J (1985) *Statistical Decision Theory and Bayesian Analysis* (Springer, New York), 2nd Ed.
19. Manski C (2004) Statistical treatment rules for heterogeneous populations. *Econometrica* 72(4):1221–1246.

20. Manski C (2005) *Social Choice with Partial Knowledge of Treatment Response* (Princeton Univ Press, Princeton).
21. Manski C (2007) Minimax-regret treatment choice with missing outcome data. *J Econom* 139(1):105–115.
22. Manski C, Tetenov A (2007) Admissible treatment rules for a risk-averse planner with experimental data on an innovation. *J Stat Plan* 137(6):1998–2010.
23. Hirano K, Porter J (2009) Asymptotics for statistical treatment rules. *Econometrica* 77(5):1683–1701.
24. Schlag K (2006) ELEVEN – Tests needed for a recommendation, EUI Working Paper ECO no. 2006/2 (European University Institute, Florence, Italy). Available at cadmus.eui.eu//handle/1814/3937. Accessed August 24, 2016.
25. Stoye J (2009) Minimax regret treatment choice with finite samples. *J Econom* 151(1):70–81.
26. Stoye J (2012) Minimax regret treatment choice with covariates or with limited validity of experiments. *J Econom* 166(1):138–156.
27. Tetenov A (2012) Statistical treatment choice based on asymmetric minimax regret criteria. *J Econom* 166(1):157–165.
28. Hróbjartsson A, Gøtzsche PC (2001) Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *N Engl J Med* 344(21):1594–1602.
29. Lichtenberg P, Heresco-Levy U, Nitzan U (2004) The ethics of the placebo in clinical practice. *J Med Ethics* 30(6):551–554.
30. Sedgwick P (2014) Clinical significance versus statistical significance. *BMJ* 348:g2130.
31. Materson BJ, et al.; The Department of Veterans Affairs Cooperative Study Group on Antihypertensive Agents (1993) Single-drug therapy for hypertension in men. A comparison of six antihypertensive agents with placebo. *N Engl J Med* 328(13):914–921.
32. Fleiss J (1973) *Statistical Methods for Rates and Proportions* (Wiley, New York).
33. Hoeffding W (1963) Probability inequalities for sums of bounded random variables. *JASA* 58(301):13–30.
34. Lugosi G (2002) Pattern classification and learning theory. *Principles of Nonparametric Learning*, ed Gyorfi L (Springer, Vienna), pp 1–56.
35. Bentkus V (2004) On Hoeffding's inequalities. *Ann Probab* 32(2):1650–1673.