Published in final edited form as:

Stat Biopharm Res. 2020 ; 12(2): 130–136. doi:10.1080/19466315.2019.1668843.

Bayesian Methods in Regulatory Science

Gary L. Rosner

Division of Oncology Biostatistics & Bioinformatics, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore MD 21205

Abstract

Regulatory science comprises the tools, standards, and approaches that regulators use to assess safety, efficacy, quality, and performance of drugs and medical devices. A major focus of regulatory science is the design and analysis of clinical trials. Clinical trials are an essential part of clinical research programs that aim to improve therapies and reduce the burden of disease. These clinical experiments help us learn about what works clinically and what does not work. The results of clinical trials support therapeutic and policy decisions. When designing clinical trials, investigators make many decisions regarding various aspects of how they will carry out the study, such as the primary objective of the study, primary and secondary endpoints, methods of analysis, sample size, etc. This paper provides a brief review of the clinical development of new treatments and argues for the use of Bayesian methods and decision theory in clinical research.

Keywords

Clinical trials; Decision theory; Study design

1 Introduction

The U.S. Food and Drug Administration (FDA) defines regulatory science as “the science of developing new tools, standards, and approaches to assess the safety, efficacy, quality, and performance of all FDA-regulated products” (F.D.A. 2018). The primary method for assessing the characteristics of such products is the clinical trial. Clinical trials are experiments that investigators carry out to learn about the safety of a new drug or device, to get preliminary data on the activity or potential benefit of the investigational agent or device, and about whether the new treatment is better (or, at least, not worse) than the current standard of care. Even though regulators consider drugs, devices, and other modalities to treat diseases, injuries, and other health-related conditions, our discussion considers the drug development paradigm as a general framework for evaluating new treatment regimens.

Drug development proceeds in phases. We may be interested in learning about which doses are safe (phase 1), if a new form of therapy shows some activity and promise for treating some disease or condition (phase 2), and whether use of the treatment leads to better outcomes for patients than one sees among similar patients who receive a current standard of

care, or which adverse events may occur with a particular drug or device (phase 3). In general, final regulatory approval will almost always rest on a demonstration of some benefit (perhaps as non-inferiority) of the treatment in well-designed randomized clinical trials. Thus, devising designs and tools for the analysis of clinical trials are key concerns among regulatory scientists.

Results of clinical trials aid medical decision-making. The decisions might be which doses to administer, how often the patients should take the medication, whether to treat a patient with a device or a drug, or whether to treat any patients with the new therapy at all. Decision making is nowhere more evident than in the choice of which treatment to recommend or choose. No matter the endpoint, a well-motivated and appropriately designed study will inform clinical decision-making.

A clinical trial requires many decisions during its planning and throughout the course of the study. Investigators may have to decide on the study's primary objective, endpoints, the number of patients to enroll, the appropriate control therapy, and the inclusion of interim analyses, to name a few. Given the large number of decisions that must take place when planning and running a clinical study, it seems prudent to incorporate decision making into the study design process in a formal way. This paper reviews development of medical products and presents some reasons why investigators should consider incorporating formal decision theory in clinical trial design and why a Bayesian decision-theoretic approach may lead to better outcomes than current reliance on an amalgamation of the Neyman-Pearson hypothesis testing framework and p values. Manski (2019) argues for the use of decision theory in clinical trials using frequentist approaches.

2 Background

2.1 Bayesian clinical trials

It is worth pointing out that there are several different types of study designs that people have called Bayesian designs. Broadly, there are two categories: those one might call truly Bayes and those that are stylized or calibrated Bayes (Little 2006). A subset of the first group consists of Bayesian optimal designs. When designing a study, a Bayesian will choose design points to maximize the expected utility. Here, expectation is with respect to our current state of knowledge. That is, the design includes consideration of all sorts of previous and current information, and it does so via mathematical characterization of this information and probability calculus.

Stylized Bayes designs, on the other hand, use Bayesian models to characterize the study data but adjust design parameters to achieve desirable frequentist operating characteristics. Many such designs treat the parameters in the prior distributions as design parameters and adjust these distributional parameters or certainty thresholds in the service of satisfying perceived frequentist desiderata.

One area of regulatory science in which a Bayesian approach of either category has found acceptance (or, at least, an appreciation for this approach) is when one wants to borrow information. An example is when a medical device manufacturer makes a small change to an

existing device (Campbell 2011). Bayesian modeling has also found some acceptance in population pharmacokinetics and pharmacodynamics, because of the use of hierarchical (or population) models in these areas of research (Davidian & Giltinan 1995). There are also proposals for Bayesian modeling in pediatric drug development if one wants to borrow information available from studies of the same therapy in adults (Schoenfeld et al. 2009, Gamalo-Siebers et al. 2017). One may also want to borrow information when evaluating treatments for rare diseases or when one is concerned about rare adverse treatment side effects. Bayesian methods also feature prominently in many newer designs for early phase clinical studies, particularly for phase 1 and 2 studies (Spiegelhalter et al. 2004, Berry et al. 2011, Yuan et al. 2016).

One broad way to view development of new therapies in regulatory science is as a series of learn-confirm cycles, as discussed by Sheiner (1997). Sheiner proposed that there are two cycles. One learn-confirm cycle consists of phase 1 and some phase 2 studies. The second cycle involves a so-called phase 2b study and phase 3 studies. These categories are not 100% of one type or the other. Phase 1 studies learn about dosing and which doses or ways of administering the therapy patients find tolerable. Phase 2 studies seek to confirm the activity (or the promise) of the therapy under evaluation. At the end of this cycle, investigators make a decision about whether to carry out further evaluation of the treatment. If the decision is to evaluate the treatment further, then the second learn-confirm cycle commences. This second cycle consists of phase 2B studies, in which one learns how to use the therapy for maximum effect. These studies are followed by phase 3 studies to confirm that the therapy is efficacious and has an acceptable benefit-risk ratio in a large population of patients who represent individuals for whom the therapy is intended.

This cycling between learning and confirming treats each part as distinct. The goals of the learning and confirming stages are different, however. The design and methods of analysis will depend on whether the goal is to learn or to confirm. Learning involves estimation, whereas confirming involves hypothesis testing in Sheiner's view (Sheiner 1997). Bayesian methods lend themselves clearly to estimation by virtue of being built on probability theory to provide a coherent system for learning via updating knowledge with new information. The Bayesian approach mathematically provides a natural progression from prior uncertainty to diminished posterior uncertainty after accounting for information in the study data. Sheiner considers hypothesis testing as the inferential framework for the confirmatory part of the learn-confirm cycle. Hypothesis tests seek to answer the question of whether observations agree with or conflict with predictions based on specific hypotheses.

2.2 Hypothesis testing and decision making

In consideration of hypothesis testing, Neyman and Pearson formulated a rule that limits the risks of committing certain errors in the long run. In this framework, one chooses between two competing hypotheses, H_0 and H_1 , each reflecting a different state of nature. As Neyman & Pearson (1933) wrote, "Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong." Neyman and Pearson considered the choice of hypothesis as a decision

problem. The possible actions when choosing between two hypotheses are to accept one or to reject it in favor of the other hypothesis. In a decision-theoretic formulation, one will incur a loss if one accepts a false hypothesis or if one rejects a true hypothesis.

The Neyman-Pearson hypothesis-testing framework limits the risk of incorrect decisions from among a multitude of decisions regarding statistical hypotheses, focusing on the long-term risks of making errors regarding null and alternative hypotheses (Neyman & Pearson 1928a,b, 1933). This framework does not lend itself to (nor was it intended to) making statements beyond, “We reject H_0 in favor of H_1 ” for the study at hand. The resulting decision rule does not tell us whether we are reaching the wrong decision in this study by rejecting or failing to reject based on this study’s data or any other information. Instead, the decision rule places inference for this study within the context of inferences among all studies one might encounter.

The analysis of a clinical study usually includes a statement about the decision to reject the null hypothesis and a p value to accompany a corresponding analysis. The trialists thus assess the support for the null hypothesis in a study’s data by computing a measure of significance that is attributable to Fisher (1954). In other words, the authors make a decision about the statistical hypotheses based on the trial data and try to validate the decision with the p value as a measure of the strength of evidence that supports the decision. Adding a p value does not help, though. A p value does not provide a measure of evidence in favor of anything, since its use in hypothesis testing is akin to proof by contradiction. This inferential framework only tells us something about the null hypothesis’s support for the observed data, not the support in the data for either hypothesis. In medical journals, we find the use of statistical significance as a measure of support for any conclusion from the trial, as well as a basis for making decisions. Many of these journals reserve the term “significant” for statements for which a statistical hypothesis test yields a p value below some threshold, typically 0.05. (We note that this practice may be changing in response to the American Statistical Association’s Statement on P values (Wasserstein & Lazar 2016). Among medical journals, for example, see Harrington et al. (2019)).

Current application of the frequentist inferential paradigm in biomedical research is an amalgamation of the two distinct approaches to inference: hypothesis testing and p values as measures of evidence. On the one hand, trial designs seek to limit the risks of Type 1 or Type 2 errors (Neyman-Pearson hypothesis testing). The decision whether to reject the statistical null hypothesis will compare a test statistic to a predetermined threshold that limits the long-run probability of a Type 1 error. Any considerations relating to optimizing the number of patients who benefit from the treatment or minimizing costs are usually considered only informally.

As pointed out by Goodman (1999a,b), this practice of using p values within the Neyman-Pearson testing framework is logically flawed, because it conflates two distinct and conflicting ideas. Whereas statistical significance says something about the consistency of the current study’s data and a particular hypothesized state of nature, it does not help us ensure that we limit the risk of reaching the wrong conclusion in the long run. Regulators, however, need to make decisions about specific treatments, devices, etc., based on the

totality of evidence, including what is known about alternative treatment options and predicted safety and benefit. Neither the Neyman-Pearson framework nor a p value seem to provide a firm basis for this task.

2.3 Early considerations of the use of decision theory in clinical research

An early call for the use of decision theory in clinical studies was contained in Anscombe's review (Anscombe 1963) of the Armitage book *Sequential Medical Trials* (Armitage 1960). Armitage's book proposed designs for clinical experiments that built on the methodologic research into minimizing destructive testing that occurred during World War 2 and led to the sequential tests of Wald (1945) and Barnard (1946). The main motivation for introducing sequential designs in clinical trials was to shorten the time until a study can reach a conclusion and the investigators can publish the trial's results. Anscombe's criticisms relate to the proposed statistical framework's confusion of the treatment decisions that physicians may make as a result of the study and the presentation of the study results with respect to the treatment comparison.

Additionally, Anscombe points out that a Neyman-Pearson approach to hypothesis testing may make sense in some areas but not in scientific experimentation. For example, in the context of quality control one wishes to ensure long-term proper operations of a manufacturing facility or profitability of the company. In such situations, decision rules like that of Neyman and Pearson make some sense, since the rule maintains the long-run properties of the process. Medical experiments are different, however. In medical research, the goal is to evaluate a specific treatment and determine if it might improve outcomes for patients relative to a comparator. While one may have to make decisions about multiple treatments based on a multitude of trials, one does not expect the same clinical trial to be repeated over and over again.

Anscombe states that there are two different phases to scientific experiments: (1) design and (2) presentation of results. While both phases involve decisions, applying the notion of routine decisions, such as those one might consider in an industrial setting, does not seem appropriate for either phase of scientific experiment. In particular, he argues for presentation of the study data (principally via the likelihood function) as the primary responsibility of the investigators. Neyman-Pearson hypothesis testing and p values do not seem to him to play any role here. We are specifically interested in this experiment's results when we read about the study and not those outcomes that might have occurred but did not.

Aside from the use of a study's data for making decisions relating to policy, regulation, or therapy, decision theory has a role to play when designing a clinical study. Around the same time that Anscombe's review appeared, Colton (1963) proposed a study design with a stopping boundary based on the number of future patients who might be affected by the treatment under evaluation. (Anscombe presents a similar idea in his review.) With improved and more efficient algorithms for easing the computational burden associated with finding (or approximating) the optimal Bayesian design in a decision-theoretic way, interest in applying this approach to designing studies appears to have grown over the past thirty years.

2.4 Bayesian hypothesis testing

The Bayesian inferential paradigm does provide the necessary tools with which to make probability statements about a study, a treatment, or a hypothesis. The paradigm lends itself to learning. Starting from prior uncertainty about a treatment, for example, we carry out an experiment to collect data and learn about the treatment's effects in the study population. The experiment should reduce our uncertainty and increase our knowledge about the treatment. The application of probability calculus allows us to revise our prior uncertainty in light of the study data, leading to less uncertainty. Bayesian statistics allows for more than just combining the probabilistic characterization of the study data (i.e., the likelihood or sampling distribution) with the mathematical characterization of prior uncertainty (the prior distribution) to produce the posterior distribution as an updated characterization of uncertainty. The mathematical and statistical models a Bayesian uses lend themselves to inclusion of information, data, and beliefs or opinions that are external to the study. With Bayes rule, the many sources of uncertainty and heterogeneity combine in accordance with the laws of probability, allowing an assessment of the final inference's uncertainty.

Bayesian hypothesis testing seems to be underutilized or, perhaps, less well known. Bayesian hypothesis tests evaluate the support in the data for any number of hypotheses under consideration. Thus, even if one accepts that confirmatory studies require hypothesis testing, Bayesians have something to contribute. One might base a test on a comparison of posterior probabilities, such as $\Pr(H_0|\text{Data})$ to $\Pr(H_1|\text{Data})$, or on Bayes factors, $p(\text{Data}|H_0)/p(\text{Data}|H_1)$. Bayesian hypothesis testing based on Bayes factors, however, seems to be better suited for confirmatory studies than p values (Kass & Raftery 1995, Johnson 2005, Johnson & Cook 2009). Bayes factors actually provide a measure of the evidence in a study's outcome data in favor of one hypothesis over an alternative hypothesis.

A Bayesian hypothesis test may focus solely on the current study's data, which is the current practice of applying frequentist hypothesis tests. Bayesian inference, however, also allows one to incorporate previous studies or other information in a formal way, whether through estimation or Bayesian hypothesis testing. We can make probabilistic statements about the current study in the context of what else we know that bears on the question at hand. What does this study tell us? How do these data increase our knowledge? By considering the current study within the larger context of what we know, we can quantify or, at least, characterize any change in certainty or knowledge thanks to this study. Bayesian inference provides a framework for interpreting the current study's results in light of everything else we know; a significance test does not.

3 Decision theory and clinical research

3.1 Purpose of a clinical trial

What do we want from a clinical trial? Do we want a measure of the evidence that one treatment is superior to the other? Do we want an estimate of a treatment's clinical benefit, along with a measure of the precision of this estimate? Perhaps we want a decision rule regarding the hypothesis, such as "Yes, we reject the hypothesis that the treatments are equivalent." Or, do we want the current trial to inform our decision regarding the next step in

the process of developing the treatment? That is, should we continue studying the treatment, should we approve the treatment for some indication, or should we put aside any further consideration of the treatment?

Most of these possible uses relate to making decisions. That is, in most cases we want the clinical trial to inform decision-making. The decisions could be which of several treatment options to recommend to patients with a particular disease, which dose to prescribe to patients with a particular condition, or whether to consider further clinical evaluation of a new therapeutic approach. If the purpose of running the trial is to help us make decisions, then we should consider applying decision theory in a formal way when determining the import of the trial's results. That is, we want to consider the clinical trial in the context of overall medical decision-making.

We also have to make many decisions when we are designing a clinical trial. We need to determine how many patients to enroll in the study, what the study's primary endpoint should be, how long we need to follow these patients based on the chosen endpoint, how often we want to carry out interim analyses and apply stopping rules, and, most of all, whether to carry out the study in the first place. Why not make these decisions explicitly part of the design of the study? If we are going to apply mathematical rigor to these decisions, why not do so in a coherent way? Statistical decision theory provides the tools to make coherent decisions, and Bayes rules have optimality properties relating to decision-making.

3.2 Decision theory in the design of clinical trials

Many people use the posterior distribution for decision making, since it contains all "knowledge" at the time of the decision. Aside from providing probability statements about the truth of hypotheses or likely values of treatment effects, however, Bayesian calculations also allow one to make probabilistic statements about future events. Not only can we provide a current assessment of the probability that a null or alternative hypothesis is true—something lacking from the Neyman-Pearson approach. We can also determine the probability that a future patient will benefit from the treatment without suffering serious side effects or that the next study will show similar results, given the current study and all currently available information. Optimal Bayesian decisions maximize the expected utility of the decision maker over all possible decisions (Lindley 1985, Raiffa & Schlaifer 2000).

We consider each decision an action. Determining the outcome and expected utility of taking some action requires that one predict future outcomes that may arise if one takes the action, along with the likelihood of the outcomes. Therefore, we believe that the predictive distribution is very much a key element of optimal Bayesian decision making. Of course, one needs the posterior distribution (or prior if making a decision ahead of incorporating study data) to compute the predictive distribution, so the posterior distribution is vital to optimal decision making, too. The Bayesian framework provides a means to characterize the probabilities of as yet unseen observations and order them from most likely to least likely, conditional on the data and any other information that one already has in hand. The calculation of the predictive distribution accounts for uncertainty in model parameters, as characterized by the posterior distribution, and variation in the data, allowing for a coherent incorporation of uncertainty in predictions.

An optimal decision from a Bayesian perspective will be the decision that maximizes the expected utility. One determines the optimal decision by calculating (or approximating) the expected utility for all possible decisions, accounting for uncertainty in the statistical model, including uncertainty in predictions of future observations. Computation of the predictive distribution is the means by which the uncertainties and heterogeneities in the data and other sources of information propagate in a manner that is consistent with probability calculus to characterize the uncertainty about future outcomes.

If one wants to apply decision theory in the context of a clinical trial, one will consider a set of possible actions, $a \in \mathcal{A}$, a sampling distribution to characterize variation in possible outcomes among a sample of patients, $\mathbf{Y}$, and the parameters, Θ , that characterize the sampling distribution. Additionally, one has to bring together the costs and benefits in a utility function, $u(\cdot)$. The utility function provides a means to quantify the relative merits or costs of outcomes arising as a result of decisions or actions. The function might include elements, such as the number of patients, study duration, or the fraction of responding patients. For a given action, the expected utility integrates all uncertainty in the data and unknown parameters: $\mathcal{U}(a) = \int_{\mathbf{Y}} \int_{\Theta} u(y) p_a(y|\theta) p(\theta) d\theta dy$. A Bayesian decision rule chooses the action that maximizes the expected utility, namely $a^* = \arg \max \mathcal{U}(a)$.

The utility function can contain both losses and benefits. A benefit term might also be the potential gain if a future study achieves a statistically significant result. For example, Ding et al. (2008) consider the following utility function to aid decision making at interim and the final analysis of a phase 2 study. The interim decisions may be one of three actions: $a = 1$ corresponds to enrolling another cohort of patients into the study; $a = 2$ is the decision to stop the study and abandon the treatment from further evaluation; and $a = 3$ corresponds to embarking on a future randomized phase 3 study to compare the treatment to a standard of care. The utility function is given by

$u(a) = -\{(c_1 \times n_1 \times t) + (c_2 \times n_2)\} + \{b \times (\theta_{new} - \theta_{old}) I[z > z_1 - \alpha]\}$ (Ding et al. 2008). The function considers the per-patient cost of enrolling patients (c_1) in t cohorts of n_1 patients in the phase 2 study, as well as the cost (c_2) of enrolling n_2 patients in a future phase 3 study. In our experience, incorporating a future hypothesis test in the utility function, along with other important study considerations, leads to a study with good frequentist operating characteristics, as well as a study that achieves the goals of its organizers.

The general idea is to determine a utility function that incorporates potential gains and losses. For each possible action, one finds the expected utility by integrating the utility function with respect to possible variation in patient outcomes and uncertainty in model parameters. Statistical decision theory recommends that one choose the action with the largest expected utility. Stallard et al. (1999) discuss decision-theoretic designs for phase 2 studies, for example. Lewis et al. (2007) also advocate strongly for group sequential designs that incorporate decision theory in their designs. Müller et al. (2017) describe a particular application of decision theory in the design of a clinical trial for which one includes a Bayesian nonparametric probability model.

Several authors have discussed the application of decision-theoretic considerations when choosing a sample size for a study, for example Bernardo (1997) and De Santis & Gubbiotti

(2017). Some authors focus on the benefits of decision theory in study design by basing the utility function on some measure of information (Sebastiani & Wynn 2000, Ventz et al. 2018), particularly as it relates to the size of the population (Cheng et al. 2003, Stallard et al. 2017). Other examples include the determination of sampling times for a complex pharmacokinetics study (Stroud et al. 2001) and the design of a bioequivalence study (Lindley 1998). In the context of studies that seek the right doses or schedules for therapies, several researchers have proposed designs that ask the investigators to assign weights to various adverse events. For example, Thall & Cook (2004) suggest that one consider the benefits and risks when one seeks the right dose or doses of drugs to administer to patients. Ezzalfani et al. (2013) propose combining adverse events into a continuous score based on relative weights of the side effect. Each of these examples has the advantage that clinicians and, potentially, patients may start to think about treatment choices more rationally than before. Table 1 lists several areas in which applying Bayesian decision theory in the design of clinical trials may mitigate some concerns with the current prevailing approach to designing clinical trials.

A challenge one faces when applying decision theory to design a clinical study is the formation of a utility function. Some outcomes that investigators wish to include in the function are on different scales. For example, there is the cost of enrolling patients, screening patients for eligibility, and following patients, which can be measured in monetary units. A utility function may also include the precision of estimation, the general benefit of a clinical response, extending time without disease, reducing the risk of serious adverse events, or some measure relating to the inconvenience caused by making patients wait in the clinic for post-treatment evaluations. Each of these consequences of a study's design is measured on different scales. One can sometimes assign monetary values to such non-fiscal outcomes, but some attributions are not often straightforward. Some investigations have considered endpoints such as quality-adjusted life years that combine good and bad clinical outcomes corresponding to different health states (Weinstein et al. 2009, Devlin & Lorgelly 2017). Eliciting utilities from clinicians or patients for use in decision making has also been the focus of research, for example, in the evaluation of treatment choices for head and neck cancers (Ramaekers et al. 2011, Meregaglia & Cairns 2017). One approach we have used assigns weights to the constituent components of the utility function such that the weights are all multiples of one of the elements, such as a particular outcome. By varying the multiplication factors of one or several of these individual relative weights, one can examine the sensitivity of the optimal design and various study characteristics to these multiplicative constants or weights. While accounting for potential outcomes and assigning utilities to these are challenging tasks, they are not insurmountable; in fact, we believe that the exercise is often illuminating.

A particularly vexing problem with determining a fully sequential design is the computational burden. Fully sequential trial designs require one to use computationally complex algorithms, such as backward induction (Brockwell & Kadane 2003, Müller et al. 2007), which can be prohibitively difficult. Methods for finding approximately optimal fully-sequential designs exist, however, and the literature includes some examples in which these approximations have been applied (Carlin et al. 1998, Ding et al. 2008, Rossell et al. 2007).

4 Conclusion

The emphasis in clinical research should be on designing studies appropriately and presenting the study results transparently and without bias. Decision theory has a role to play in the design of efficient studies that account for the needs of various stakeholders. Fair and accurate presentations of outcomes of well-designed studies will provide regulators, caregivers, and patients the information they need to make the decisions that are the right ones for them. These are principles on which we should be able to agree, even if we do not all share the opinion that the Bayesian inferential paradigm provides the right framework for carrying out clinical research. Decision theory provides a basis for designing studies in a manner that makes clear the relative importance the study investigators have placed on certain aspects of the treatment or potential outcomes of the study. Examining a utility function and its components with investigators during the design stage and including a discussion of the utility function when presenting the study's results will promote transparency to the broader community.

Although there are non-Bayesian approaches to decision theory, the Bayesian paradigm lends itself to optimal decision making in a natural way. Acceptance of Bayesian approaches, however, has lagged in many situations relating to clinical research. Many still argue that Bayesian inference is too subjective, and the critics argue that one should be "objective" when designing and carrying out clinical studies. These critics express concern about the prior distribution exerting too much influence on the inference. These criticisms have a very long history. It is well known, however, that the frequentist approach also includes much subjectivity, such as the choice of null and alternative hypotheses, translation of p values into evidence from a study, and, most notably, how one incorporates any one study in the overall context of the study's subject matter (Berger & Berry 1988).

Interest in understanding and applying statistical decision theory and Bayesian inference has grown among our medical research colleagues (Henriquez & Korpi-Steiner 2016, Operskalski & Barbey 2016). We statisticians are in a unique position to answer calls for help from the biomedical community. Aside from helping to align the goals of clinical trials with the designs, statisticians applying statistical decision theory in the service of clinical research also open up important research opportunities.

In summary, Bayesian inference, particularly Bayesian decision theory, has a place in regulatory science. The learn-confirm cycle provides a paradigm within which Bayesian methods and decision theory clearly each have a place. Framing many study design considerations as decisions, including those relating to the study's primary objective and final inference, has the potential to allow clinical studies to better address the needs of the studies' many stakeholders. Such approaches may provide patients and their caregivers better tools with which to make informed decisions.

Acknowledgments

The author gratefully acknowledges constructive comments from three reviewers and an associate editor that improved the manuscript. Some of this work was supported by grant *NCI P30CA006973* and by a Center of Excellence in Regulatory Science and Innovation (CERSI) grant to Johns Hopkins University from the US Food &

Drug Administration U01FD005942. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the HHS or FDA.

References

- Anscombe FJ (1963), 'Sequential medical trials', *Journal of the American Statistical Association* 58, 365–383.
- Armitage P (1960), *Sequential Medical Trials*, Thomas, Springfield, Ill.
- Barnard GA (1946), 'Sequential tests in industrial statistics', *Journal of the Royal Statistical Society Series B-Statistical Methodology* 8(1), 1–26.
- Berger JO and Berry DA (1988), 'Statistical analysis and the illusion of objectivity', *American Scientist* 76(2), 159–165.
- Bernardo JM (1997), 'Statistical inference as a decision problem: The choice of sample size', *Statistician* 46(2), 151–153.
- Berry SM, Carlin BP, Lee JJ and Müller P (2011), *Bayesian Adaptive Methods for Clinical Trials*, CRC Press, Boca Raton.
- Brockwell AE and Kadane JB (2003), 'A gridding method for Bayesian sequential decision problems', *Journal of Computational and Graphical Statistics* 12, 566–584.
- Campbell G (2011), 'Bayesian statistics in medical devices: Innovation sparked by the FDA', *Journal of Biopharmaceutical Statistics* 21(5), 871–887. [PubMed: 21830920]
- Carlin BP, Kadane JB and Gelfand AE (1998), 'Approaches for optimal sequential decision analysis in clinical trials', *Biometrics*. 54(3), 964–75. [PubMed: 9750245]
- Cheng Y, Su F and Berry DA (2003), 'Choosing sample size for a clinical trial using decision analysis', *Biometrika* 90(4), 923–936.
- Colton T (1963), 'A model for selecting one of two medical treatments', *Journal of the American Statistical Association* 58(302), 388–400.
- Davidian M and Giltinan D (1995), *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall, London.
- De Santis F and Gubbiotti S (2017), 'A decision-theoretic approach to sample size determination under several priors', *Applied Stochastic Models in Business and Industry* 33(3), 282–295.
- Devlin NJ and Lorgelly PK (2017), 'QALYs as a measure of value in cancer', *Journal of Cancer Policy* 11, 19–25. *Value and Cancer*.
- Ding M, Rosner GL and Mueller P (2008), 'Bayesian optimal design for phase II screening trials', *Biometrics* 64(3), 886–94. [PubMed: 18162113]
- Ezzalfani M, Zohar S, Qin R, Mandrekar SJ and Deley M-CL (2013), 'Dose-finding designs using a novel quasi-continuous endpoint for multiple toxicities', *Statistics in Medicine* 32(16), 2728–2746. [PubMed: 23335156]
- F.D.A. (2018), 'Advancing regulatory science', <https://www.fda.gov/ScienceResearch/SpecialTopics/RegulatoryScience/default.htm>. Accessed: June 5, 2018.
- Fisher RA (1954), *Statistical Methods for Research Workers*, 12th ed., rev. edn, Oliver and Boyd, Edinburgh.
- Gamalo-Siebers M, Savic J, Basu C, Zhao X, Gopalakrishnan M, Gao A, Song G, Baygani S, Thompson L, Xia HA, Price K, Tiwari R and Carlin BP (2017), 'Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation', *Pharmaceutical Statistics* 16(4), 232–249. [PubMed: 28448684]
- Goodman SN (1999 a), 'Toward evidence-based medical statistics. 1: The p value fallacy', *Annals of Internal Medicine* 130(12), 995–1004. [PubMed: 10383371]
- Goodman SN (1999 b), 'Toward evidence-based medical statistics. 2: The Bayes factor', *Annals of Internal Medicine* 130(12), 1005–13. [PubMed: 10383350]
- Harrington D, D'Agostino RB, Gatsonis C, Hogan JW, Hunter DJ, Normand S-LT, Drazen JM and Hamel MB (2019), 'New guidelines for statistical reporting in the journal', *New England Journal of Medicine* 381(3), 285–286. [PubMed: 31314974]

- Henriquez RR and Korpi-Steiner N (2016), 'Bayesian inference dilemma in medical decision-making: A need for user-friendly probabilistic reasoning tools', *The Clinical Chemist* 62(9), 1285–1286.
- Johnson VE (2005), 'Bayes factors based on test statistics', *Journal of the Royal Statistical Society Series B-Statistical Methodology* 67, 689–701.
- Johnson VE and Cook JD (2009), 'Bayesian design of single-arm phase II clinical trials with continuous monitoring', *Clinical Trials* 6(3), 217–226. [PubMed: 19528131]
- Kass RE and Raftery AE (1995), 'Bayes factors', *Journal of the American Statistical Association* 90(430), 773–795.
- Lewis RJ, Lipsky AM and Berry DA (2007), 'Bayesian decision-theoretic group sequential clinical trial design based on a quadratic loss function: A frequentist evaluation', *Clinical Trials* 4(1), 5–14. [PubMed: 17327241]
- Lindley DV (1985), *Making Decisions*, 2nd edn, John Wiley & Sons Ltd, London.
- Lindley DV (1998), 'Decision analysis and bioequivalence trials', *Statistical Science* 13(2), 136–141.
- Little RJ (2006), 'Calibrated Bayes: A Bayes/frequentist roadmap', *The American Statistician* 60(3), 213–223.
- Manski CF (2019), 'Treatment choice with trial data: Statistical decision theory should supplant hypothesis testing', *The American Statistician* 73(sup1), 296–304.
- Meregaglia M and Cairns J (2017), 'A systematic literature review of health state utility values in head and neck cancer', *Health and Quality of Life Outcomes* 15(1), 174. [PubMed: 28865475]
- Müller P, Berry DA, Grieve AP, Smith M and Krams M (2007), 'Simulation-based sequential Bayesian design', *Journal of Statistical Planning and Inference* 137(10), 3140–3150.
- Müller P, Xu Y and Thall PF (2017), 'Clinical trial design as a decision problem', *Applied Stochastic Models in Business and Industry* 33(3), 296–301. [PubMed: 29200977]
- Neyman J and Pearson E (1933), 'On the problem of the most efficient tests of statistical hypotheses', *Philosophical Transactions of the Royal Society of London, Series A* 231, 289–337.
- Neyman J and Pearson ES (1928 a), 'On the use and interpretation of certain test criteria for purposes of statistical inference. Part I', *Biometrika* 20A, 175–240.
- Neyman J and Pearson ES (1928 b), 'On the use and interpretation of certain test criteria for purposes of statistical inference. Part II', *Biometrika* 20A, 263–294.
- Operskalski JT and Barbey AK (2016), 'Risk literacy in medical decision-making', *Science* 352(6284), 413–414. [PubMed: 27102467]
- Raiffa H and Schlaifer R (2000), *Applied Statistical Decision Theory*, Wiley Classics Library, Wiley.
- Ramaekers BL, Joore MA, Grutters JP, van den Ende P, de Jong J, Houben R, Lambin P, Christianen M, Beetz I, Pijls-Johannesma M and Langendijk JA (2011), 'The impact of late treatment-toxicity on generic health-related quality of life in head and neck cancer patients after radiotherapy', *Oral Oncology* 47(8), 768–774. [PubMed: 21683647]
- Rossell D, Mueller P and Rosner GL (2007), 'Screening designs for drug development', *Biostatistics* 8(3), 595–608. [PubMed: 17030589]
- Schoenfeld DA, Hui Z and Finkelstein DM (2009), 'Bayesian design using adult data to augment pediatric trials', *Clinical Trials* 6(4), 297–304. [PubMed: 19667026]
- Sebastiani P and Wynn HP (2000), 'Maximum entropy sampling and optimal Bayesian experimental design', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62(1), 145–157.
- Sheiner LB (1997), 'Learning versus confirming in clinical drug development', *Clinical Pharmacology and Therapeutics* 61(3), 275–291. [PubMed: 9084453]
- Spiegelhalter DJ, Abrams RK and Myles JP (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, John Wiley & Sons, Chichester, UK.
- Stallard N, Miller F, Day S, Hee SW, Madan J, Zohar S and Posch M (2017), 'Determination of the optimal sample size for a clinical trial accounting for the population size', *Biometrical Journal* 59(4), 609–625. [PubMed: 27184938]
- Stallard N, Thall P and Whitehead J (1999), 'Decision theoretic designs for phase II clinical trials with multiple outcomes', *Biometrics* 55, 971–977. [PubMed: 11315037]

- Stroud JR, Müller P and Rosner GL (2001), 'Optimal sampling times in population pharmacokinetic studies', *Journal of the Royal Statistical Society Series C-Applied Statistics* 50(3), 345–359.
- Thall PF and Cook JD (2004), 'Dose-finding based on efficacy-toxicity trade-offs', *Biometrics* 60(3), 684–693. [PubMed: 15339291]
- Ventz S, Cellamare M, Bacallado S and Trippa L (2018), 'Bayesian uncertainty directed trial designs', *Journal of the American Statistical Association* 0(0), 1–13.
- Wald A (1945), 'Sequential tests of statistical hypotheses', *Annals of Mathematical Statistics* 16(2), 117–186.
- Wasserstein RL and Lazar NA (2016), 'The ASA statement on p-values: Context, process, and purpose', *The American Statistician* 70(2), 129–133.
- Weinstein MC, Torrance G and McGuire A (2009), 'QALYs: The basics', *Value in Health* 12, S5–S9. [PubMed: 19250132]
- Yuan Y, Nguyen HQ and Thall PF (2016), *Bayesian Designs for Phase I-II Clinical Trials*, Chapman & Hall, Boca Raton.

Table 1

Concerns with the current paradigm for evaluating treatments that the application of decision theory could mitigate.

Current paradigm	Bayesian decision-theoretic design
Emphasizes statistical significance over clinical importance	Incorporates measures of clinical utility (benefits and risks)
Treats each study as if no other studies exist, thereby ignoring external information in the formal analysis of the study's data	Allows formal incorporation of multiple sources of relevant knowledge in a probabilistically coherent way via Bayesian decision theory
Formally considers just one primary clinical endpoint, even though clinical usefulness of the treatment depends on many characteristics of the treatment (e.g., non-inferiority rather than benefit versus risk)	Can consider clinical benefit and risk of adverse events via utility functions, possibly allowing patient-specific weighting of potential outcomes (good and bad)
Emphasizes individual comparisons of outcomes, making it difficult for patients to interpret the data in a manner that will allow them to apply their own utilities to make their own decisions	Allows transparency of the elements of the utility functions used in the conduct of the study, making it easier for patients to relate their personal views to the study's design and results
Current paradigm	Bayesian decision-theoretic design
Does not formally consider the benefit of the treatment to patients outside of the current trial	Allows characterization of predicted outcomes and associated uncertainty via Bayesian calculations
Does not base decisions about whether or not to carry out a study on formal probability-based predictions of the future study's success	Can provide a formal probabilistic characterization of the likelihood of success of the next study evaluating the therapy
Does not update knowledge in a formal way as the study's data accrue	Will automatically update knowledge coherently via Bayesian calculation and the posterior distribution of a treatment's effect
Has led to many studies—particularly early phase studies—that are not replicated	Could include the predictive probability of replication, leading to a more rigorous quantification of the uncertainty associated with the result and the chance a new study would replicate the inference from the current study
Does not formally consider the magnitude of losses for incorrect decisions	Incorporates weights in the utility function that reflect relative importance of outcomes