Journal of Econometrics 211 (2019) 117–136

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Is the FDA too conservative or too aggressive?: A Bayesian decision analysis of clinical trial design☆

Leah Isakov [a], Andrew W. Lo [b,*], Vahid Montazerhodjat [c,*]

[a] Sudbury Biostat, United States
[b] MIT Sloan School of Management, MIT Laboratory for Financial Engineering, MIT Computer Science and Artificial Intelligence Laboratory, 100 Main Street, E62–618, Cambridge, MA 02142, United States
[c] Department of Computer Science, Boston College, St. Mary's Hall S256, Chestnut Hill, MA 02467, United States

ARTICLE INFO

Article history:
Available online 4 January 2019

Keywords:
Clinical trial design
Drug-approval process
FDA
Bayesian decision analysis
Adaptive design

ABSTRACT

Implicit in the drug-approval process is a host of decisions—target patient population, control group, primary endpoint, sample size, follow-up period, etc.—all of which determine the trade-off between Type I and Type II error. We explore the application of Bayesian decision analysis (BDA) to minimize the expected cost of drug approval, where the relative costs of the two types of errors are calibrated using U.S. Burden of Disease Study 2010 data. The results for conventional fixed-sample randomized clinical-trial designs suggest that for terminal illnesses with no existing therapies such as pancreatic cancer, the standard threshold of 2.5% is substantially more conservative than the BDA-optimal threshold of 23.9% to 27.8%. For relatively less deadly conditions such as prostate cancer, 2.5% is more risk-tolerant or aggressive than the BDA-optimal threshold of 1.2% to 1.5%. We compute BDA-optimal sizes for 25 of the most lethal diseases and show how a BDA-informed approval process can incorporate all stakeholders' views in a systematic, transparent, internally consistent, and repeatable manner.

1. Introduction

Randomized clinical trials (RCTs) are widely accepted as the most reliable approach for determining the safety and efficacy of drugs and medical devices (Pocock, 1983; Friedman et al., 2010).[1] Their outcomes largely determine whether new therapeutics are approved by regulatory agencies such as the U.S. Food and Drug Administration (FDA). RCTs may involve a few dozen to several thousand human subjects, depending on the expected treatment effect, design choices, and business conditions, often requiring years to enroll and complete. Because of the cost and duration of these trials, the FDA is sometimes criticized for being too conservative, requiring trials that are "overly large" (Berry, 2006) and using too conservative a threshold of statistical significance.

☆ We thank Ernie Berndt, Don Berry, Bruce Chabner, Mark Davis, Hans-Georg Eichler, Williams Ettouati, Gigi Hirsch, Leonid Kogan, Tomas Philipson, Nora Yang and participants at the MIT Sloan Finance Faculty Research Retreat, the editor, Jianqing Fan, and two referees for helpful comments and discussion, and Jayna Cummings for editorial assistance. The views and opinions expressed in this article are those of the authors only and do not necessarily represent the views and opinions of any other organizations, any of their affiliates or employees, or any of the individuals acknowledged above. Research support from the MIT Laboratory for Financial Engineering is gratefully acknowledged.

* Corresponding authors.
E-mail addresses: alo-admin@mit.edu (A.W. Lo), montazer@bc.edu (V. Montazerhodjat).

[1] Although there are important differences in the approval process for drugs versus medical devices, for expositional simplicity, our use of the term "drugs" henceforth will include both types of therapeutics unless explicitly stated otherwise.

https://doi.org/10.1016/j.jeconom.2018.12.009
0304-4076/

At the heart of this debate is the legal requirement that a new drug must be safe and effective before it can be marketed to the public. Most clinical trials are designed primarily to address efficacy, while additional questions about safety and which endpoints are appropriate require much larger trials (for example, safety concerns are typically addressed by examining adverse reactions from multiple clinical programs (U.S. Food and Drug Administration, 1998)). This leads to the unavoidable regulatory tradeoff between reducing false positives (incorrectly approving an ineffective therapy) and false negatives (incorrectly rejecting an effective therapy). The probability of a false positive—called "Type I" error in the classical hypothesis testing literature—can be driven to 0% simply by not approving any drug candidates. Similarly, the probability of a false negative—called "Type II" error—can be driven to 0% by approving all drug candidates. Neither of these alternatives are acceptable; hence, the question is: what is the appropriate statistical procedure for trading off Type I versus Type II error?

The standard approach in RCTs is to specify a threshold of statistical significance that controls Type I error—typically 2.5% in each tail for a two-tailed hypothesis test or 5% for a one-tailed test. Results of RCTs with p-values less than this threshold are considered statistically significant and get approved, and those with higher p-values do not. Given this threshold, and assuming a specific magnitude for the treatment effect under the alternative hypothesis that the therapy is effective, the Type II error—typically set at 10% or 20%—is determined by the sample size of the RCT. While an adequately powered trial, i.e., a trial with a sufficiently large sample size, is required for scientifically sound conclusions, regulators are mostly concerned with Type I error to avoid exposing the wider population to a useless drug with potentially harmful side effects. According to the FDA, "the probability of Type II error is conventionally set at 10% to 20%; it is in the sponsor's interest to keep this figure as low as feasible especially in the case of trials that are difficult or impossible to repeat. Alternative values to the conventional levels of Type I and Type II errors may be acceptable or even preferable in some cases" (U.S. Food and Drug Administration, 1998). Nevertheless, it is difficult to determine whether regulatory agencies are being overly conservative or aggressive without explicitly specifying the burden of disease, i.e., the therapeutic costs and benefits for current and future patients.

In this article, we propose a framework for determining the optimal Type I and Type II errors for drug approval decisions that incorporates these therapeutic costs and benefits. Along the lines first proposed by Berry (1987) and Spiegelhalter et al. (1994) in the biomedical context, we identify specific costs to Type I and II errors and then compute the decision that minimizes the expected value of the cost using Bayesian decision analysis (BDA), where the expectation is calculated over both null and alternative hypotheses. In this framework, Type I and II errors may be assigned different costs, as first suggested by Anscombe (1963), Colton (1963) and Berry (1987), but we also take into account the delicate balance between the costs associated with an ineffective treatment during and after the trial. Given these costs, other population parameters, and prior probabilities, we can compute an expected cost for any fixed-sample clinical trial and minimize the expected cost over all fixed-sample tests to yield the BDA-optimal fixed-sample trial design.

The term "cost" in this context refers not only to direct financial costs, but also the consequences of incorrect decisions for all current and future patients. Complicating this process is the fact that these tradeoffs sometimes involve utilitarian conundrums in which small benefits for a large number of patients must be weighed against devastating consequences for an unfortunate few. Moreover, the relative costs (and risks) of potential outcomes are viewed quite differently by different stakeholders; patients dying of pancreatic cancer may not be as concerned about the dangerous side effects of an experimental drug as those patients with non-life-threatening conditions, or the shareholders of a publicly traded pharmaceutical company which will bear the enormous cost of wrongful death litigation. This difference is clearly echoed in the Citizens Council report published by the U.K.'s National Institute for Health and Care Excellence (NICE) (National Institute for Health and Care Excellence, 2008), and in a series of public meetings held by the FDA as part of its five-year Patient-Focused Drug Development Program, which showed a visible gap between patient perception of the risk/benefit ratio and that of the FDA (U.S. Food and Drug Administration, 2013a,b).

The concept of assigning costs to outcomes and employing cost minimization techniques to determine optimal decisions is well known (DeGroot, 1970). Our main contribution is to apply this standard framework to the drug approval process by explicitly specifying the costs of Type I and Type II errors using data on the burden of disease. This approach yields a systematic, objective, transparent, and repeatable process for making regulatory decisions that reflects differences in disease-specific parameters. In particular, reports by the FDA (U.S. Food and Drug Administration, 2013a,b) provide useful input for determining the relative cost parameters from the point of view of the patient, and even of the general public, in an objective and transparent manner. As suggested in Center for Devices and Radiological Health of the U.S. Food and Drug Administration (2010), using hard evidence to assign costs to different events, e.g., public health data, is a feasible remedy to the controversy which often surrounds Bayesian techniques, due to the subjective judgment involved in their cost-assignment process. In fact, Bayesian techniques have survived earlier controversy and are now used extensively in clinical trials for medical devices, in large part due to the support received from the FDA's Center for Devices and Radiological Health (CDRH) and the use of hard evidence in forming priors in those trials (Center for Devices and Radiological Health of the U.S. Food and Drug Administration, 2010).

The most important differences between the cost of the two types of errors may be from the patient perspective. A patient dying of a terminal illness that has no effective therapy may be willing to risk considerably higher Type I error than 2.5% if there is even a slight chance of receiving an effective therapy.[2] Using U.S. Burden of Disease Study 2010 data

[2] In fact, U.S. law now requires the FDA to take into account the severity of the disease in its decisions; see part 312, subpart E of title 21 of the Code of Federal Regulations (CFR) (U.S. Congress, 1999).

(Murray et al., 2013) to calibrate our relative cost parameters, we find that the current standards of drug approval are weighted more toward avoiding a Type I error than avoiding a Type II error. For example, the standard Type I error of 2.5% is considerably more conservative than the BDA-optimal Type I error of 23.9% to 27.8% (depending on the assumed magnitude of the treatment effect under the alternative hypothesis) for clinical trials of therapies for pancreatic cancer—a disease with a 5-year survival rate of 1% for stage IV patients (American Cancer Society estimate, last updated 9 January 2015 (American Cancer Society, 2015)). The BDA-optimal size for these clinical trials is larger than 23% across a wide range of alternative hypotheses, reflecting the fact that, for these desperate patients, the cost of trying an ineffective drug is much less than the cost of not trying an effective one. On the other hand, 2.5% is more aggressive than the BDA-optimal significance level of 1.2% to 1.5% for confirmatory clinical trials testing prostate cancer therapies. It is worth noting that the BDA-optimal size is larger not just for life-threatening cancers but also for serious non-cancer conditions, e.g., cirrhosis of the liver (optimal size of 15.3% to 17.7%) and hypertensive heart disease (optimal size of 7.6% to 9.4%).

The FDA is well aware of the plight of desperate patients and has gone to great lengths to expedite the approval process for drugs intended to treat serious conditions and rare diseases (U.S. Food and Drug Administration, 2006, 2013c),[3] starting with the Orphan Drug Act of 1983. Currently, four programs—the fast track, breakthrough therapy, accelerated approval, and priority review designations—provide faster reviews and/or use surrogate endpoints to judge efficacy. However, published descriptions (U.S. Food and Drug Administration, 2006, 2013c) do not indicate any differences in the statistical thresholds used in these programs versus the standard approval process, nor do they mention adapting these thresholds to the severity of the disease. Even under the traditional thresholds of statistical significance, both the U.S. and Europe have seen harmful drugs with severe side effects make their way into the market (Greener, 2005; Aronson, 2008; McNaughton et al., 2014; ProCon.org, 2014). Therefore, government agencies mandated to protect the public, like the FDA and the European Medicines Agency (EMA), are understandably reluctant to employ more risk-tolerant or aggressive statistical criteria to judge the efficacy of a drug. In practice, regulatory guidelines do allow for deviation from traditional phase 3 clinical trials for designated breakthrough therapies in cases of unmet medical need for severe diseases (U.S. Food and Drug Administration, 2006). Nevertheless, this is done on an *ad hoc* basis, and accelerated approval may require a post-approval commitment from the sponsor company for additional studies.

From the patient's perspective, the approval criteria in these programs may still seem far too conservative. Moreover, a large number of compounds are not eligible for special designations, and some physicians have argued that the regulatory requirements should be relaxed for drugs targeting non-cancerous life-threatening diseases, e.g., cirrhosis of the liver and hypertensive heart disease.

The need to balance these competing considerations in the drug approval process has long been recognized by clinicians, regulatory experts, and other stakeholders (Lenert et al., 1993; Eichler et al., 2009, 2013). Indeed, policymakers have acknowledged that serious diseases with relatively low prevalence require different statistical and clinical considerations to make product development feasible, as made evident by the Orphan Drug Act of 1983. It has also been recognized that these competing factors should be taken into account when designing clinical trials (Anscombe, 1963; Colton, 1963; Berry, 1987). One approach to quantify this need is to assign different costs to the different outcomes (Berry, 1987). As the noted biostatistician Donald Berry put it, "We should also focus on patient values, not just *p*-values" (Berry, 2015a).

In Section 2, we describe the shortcomings of a classical approach in designing a fixed-sample test. We then lay out the assumptions about the clinical trial to be designed and the primary response variable affected by the drug in Section 3. In Section 4, we introduce the BDA framework, which can be shown to mitigate the shortcomings of the classical approach, and then derive the BDA-optimal fixed-sample test. We apply this framework in Section 5 by estimating the parameters of the Bayesian model using the U.S. Burden of Disease Study 2010 (Murray et al., 2013). Using these estimates, we then compute the BDA-optimal tests for 25 of the top 30 leading causes of death in the U.S. in 2010, and report the results in Section 6. We discuss extensions and qualifications in Section 7, and conclude in Section 8.

2. Limitations of the classical approach

Two objectives must be met when determining the sample size and critical value for any fixed-sample RCT: (1) the chance of approving an ineffective treatment should be minimized; and (2) the chance of approving an effective drug should be maximized. The need to maximize the approval probability for an effective drug is obvious. In the classical (frequentist) approach to hypothesis testing—currently the most commonly used framework for designing clinical trials—these two objectives are pursued by controlling the probabilities of Type I and Type II errors. Type I error occurs when an ineffective drug is approved, and the likelihood of this error is usually referred to as the size, α, of the test. Type II error occurs when an effective drug is rejected, and the complement of the probability of this error, β, is defined as the power of the test.

It is clear that, for a given sample size, minimizing one of these two error probabilities is in conflict with minimizing the other. For example, the probability of a Type I error can be reduced to 0 by simply rejecting all drugs. Therefore, a balance must be struck between them. The classical approach addresses this issue by constraining the probability of Type I error to be less than a fixed value, usually $\alpha = 2.5\%$ for one-sided tests, while choosing a large enough sample size to maintain the power for the alternative hypothesis, right around another arbitrary level, usually $1-\beta = 80\%$ or 90%.

[3] See http://www.fda.gov/forpatients/approvals/fast/ucm20041766.htm.

The arbitrary nature of these values for the size and power of the test raises legitimate questions about their justification. These particular values correspond to situations that need not (and most likely do not) apply to clinical trials designed to test new drugs to treat human diseases. In fact, the design paradigm of clinical trials is heavily influenced by practices from other industries, particularly the manufacturing industry. Therefore, it is reasonable to ask if completely different industries should use the same values for the size and power of their tests. The consequences of wrongly rejecting a high quality product in quality testing are much different from the results of mistakenly rejecting an effective drug for patients with a life-threatening disease. In other words, there must be different *costs* associated with each of these incorrect rejections.

In addition to the arbitrary nature of the conventional values used for the size and power of RCTs, there is an important ethical issue in the classical design of clinical trials. The frequentist approach aims to minimize the chance of ineffective treatment after the trial, which is caused by Type I error. However, it does not take into account the ineffective treatment *during* the trial, and ignores the fact that at least half of the recruited subjects are exposed to ineffective treatment during the trial, assuming a balanced two-arm RCT (Berry, 1987, 2004). This ethical issue (along with financial and practical considerations) is a primary reason that the sample size in classical trial design is not further increased to achieve greater power. Recently there have been more novel frequentist designs for clinical trials, e.g., group sequential and adaptive tests, to decrease the average sample size in order to mitigate this ethical issue. However, one shortcoming of all these approaches is that they do not take into account the severity of the target disease.

Finally, the typical approach to the design of clinical trials does not explicitly account for the potential number of patients who will eventually be affected by the outcome of the trial. Patients suffering from the target disease may be affected positively in the case of an approved effective drug, or adversely in the case of an approved ineffective drug or a rejected effective drug. From these considerations, it is clear that the sample size of the trial should depend on the size of the population of patients who will be affected by the outcome of the trial, as suggested by Colton (1963), Cheng et al. (2003) and Berry (2004). We refer to the population affected by the outcome of the trial as the *target population* for the remainder of this article, and note that it is the same as the patient horizon originally proposed in Colton (1963) and Anscombe (1963) and later used in Cheng et al. (2003) and Berry (2004). This idea has an immediate and intuitive consequence: If the target population of a new drug comprises 100,000 individuals, its clinical trial must be larger than a trial designed for a drug with a target population of only 10,000 individuals. For example, the Cervarix GSK vaccine to prevent cervical cancer was approved based on a phase 3 clinical trial that enrolled 18,000 women in 14 countries, while Voraxaze (used for a small oncology population) was shown to be effective based on data from 22 patients in a single clinical trial (Paavonen et al., 2009).

3. A review of RCT statistics

In this section, we review the basic statistics of RCTs, and define the concepts and notation employed in this article. We begin with the design of the balanced two-arm RCT, in which the subjects are randomly assigned to either the treatment arm or the control arm, and there is an equal number of subjects in each. For simplicity, the focus is only on fixed-sample tests, where the number of subjects per arm, denoted by n, is determined prior to the trial and before collecting any observations. Furthermore, only after collecting all observations will a decision be made on whether or not to approve the drug. However, our approach is equally applicable to more sophisticated designs, including adaptive trials in which the sample sizes are dynamic and path-dependent functions of intermediate trial outcomes.

A quantitative primary endpoint is assumed for the trial.[4] For instance, the endpoint may be the level of a particular biochemical in the patient's blood, which is measured on a continuous scale and modeled as a normal random variable (Friedman et al., 2010; Jennison and Turnbull, 2010). The subjects in the treatment and control arms receive the drug and placebo, respectively, and each subject's response is assumed to be independent of all other responses. It is worth noting that if there exists a current treatment in the market for the target disease of the drug, then the existing drug is assumed to be administered to the patients in the control arm instead of a placebo. In either situation, it is natural to assume that the drug administered in the control arm is not toxic. The response variables in the treatment arm, denoted by $\{T_1, \ldots, T_n\}$, are assumed to be independent and identically distributed (IID), where $T_i \overset{\text{IID}}{\sim} \mathcal{N}(\mu_t, \sigma^2)$. Similarly, the control (placebo) arm responses, represented by $\{P_1, \ldots, P_n\}$, are assumed to be $P_i \overset{\text{IID}}{\sim} \mathcal{N}(\mu_p, \sigma^2)$, where the response variance in each arm is known and equal to σ^2. The response variance is assumed to be the same for both arms, but this assumption can easily be relaxed.

We confine our attention to superiority trials in which the drug candidate is likely to have either a positive effect or no effect (possibly with adverse side effects).[5] In such cases, the treatment effect of the drug, δ, is defined as the difference between the response means in the two arms, i.e., $\delta \triangleq \mu_t - \mu_p$. The event for which the drug is ineffective and has adverse side effects defines our null hypothesis, H_0, corresponding to $\delta = 0$ (the assumption of side effects is meant to represent a "worst-case" scenario, since ineffective drugs need not have any side effects). On the other hand, the alternative hypothesis,

[4] Throughout this article, we explicitly consider hypothesis testing for efficacy. In practice, most studies are designed to address treatment effect as the primary endpoint that drives sample size considerations. However, the same framework can easily be applied to evaluate safety concerns.

[5] Non-inferiority trials – in which a therapy is tested for similar benefits to the standard of care, but with milder side effects – also play an important role in the biopharma industry, and our framework can easily be extended to cover these cases.

H_1, represents a positive treatment effect, $\delta = \delta_0 > 0$. Therefore, a one-sided superiority test is appropriate for distinguishing between these two point hypotheses.

In a fixed-sample test with n subjects in each arm, we collect observations from the treatment and control arms, namely, $\{T_i\}_{i=1}^n$ and $\{P_i\}_{i=1}^n$, respectively, and form the following Z-statistic (sometimes referred to as the Wald statistic):

$$Z_n = \frac{\sqrt{\mathcal{I}_n}}{n} \sum_{i=1}^{n} (T_i - P_i), \tag{1}$$

where Z_n is a normal random variable, i.e., $Z_n \sim \mathcal{N}\big(\delta\sqrt{\mathcal{I}_n}, 1\big)$, and $\mathcal{I}_n = \frac{n}{2\sigma^2}$ is the so-called information in the trial (Jennison and Turnbull, 2010). The Z-statistic, Z_n, is then compared to a critical value, λ_n, and if the Z-statistic is smaller than the critical value, the null hypothesis is not rejected, denoted by $\widehat{H} = H_0$. Otherwise, the null hypothesis is rejected, represented by $\widehat{H} = H_1$:

$$Z_n \underset{\widehat{H}=H_0}{\overset{\widehat{H}=H_1}{\gtrless}} \lambda_n. \tag{2}$$

As observed in (2), the critical value used to reject the null hypothesis (or equivalently, the statistical significance level) is allowed to change with the sample size of the trial, hence the subscript n in λ_n. This lends more flexibility to the trial than the classical setting in which the significance level is exogenous and independent of the sample size.

Since a fixed-sample test is characterized completely by two parameters, namely, its sample size and critical value, as seen in (2), we denote a fixed-sample test with n subjects in each study arm and a critical value λ_n by $\texttt{fxd}(n, \lambda_n)$. It should be noted that, for simplicity, we use sample size and number of subjects per arm interchangeably throughout this work. Finally, the assumption that individual response variables are Gaussian is not necessary. Instead, as long as the assumptions of the Central Limit Theorem hold, the distribution of the Z-statistic, Z_n, in (1) follows an approximately normal distribution. Therefore, this model should be broadly applicable to a wide range of contexts.

4. Bayesian decision analysis

We propose a quantitative framework to take into explicit account the severity of the disease when determining the sample size and critical value of a fixed-sample test. We first define the costs associated with the trial given the null hypothesis, H_0, and the alternative, H_1. We then assign prior probabilities to these two hypotheses and formulate the expected cost associated with the trial. The optimal sample size and the critical value for the test are then jointly determined to minimize the expected cost of the trial. As stated earlier, the term "cost" in this article refers to the health consequences of incorrect decisions for all current and future patients, not necessarily the financial cost (we consider such costs in Appendix A.2).

Our methods are similar to Colton (1963), although the cost model used here is different from his. The authors of Cheng et al. (2003) have also investigated a similar problem; however, in addition to using a different model for the response variables, they consider a Bayesian trial where there is continuous monitoring of the data, and the Bayesian analysis of the observations is carried out during the trial. In contrast, we consider a classical fixed-sample test, where there is no Bayesian analysis of the observations or any change in the randomization of patients into the two arms, and only the design of the test is done in a Bayesian framework. Similar to the proposal in Grieve (2015), we assign different weights to the probabilities of Type I and Type II errors, and propose specific weights based on the U.S. Burden of Disease Study 2010 (Murray et al., 2013).

Cost model

The costs associated with a clinical trial can be categorized into two groups, in-trial costs and post-trial costs. In-trial costs, while independent of the final decision of the clinical trial, depend on the number of subjects recruited in the trial. Post-trial costs, on the other hand, depend solely on the final outcome of the trial, and are assumed to be independent of the number of recruited patients. In particular, we assume there is no post-trial cost associated with making a correct decision, i.e., rejecting an ineffective drug or approving an effective drug. We further allow asymmetric post-trial costs associated with Type I and Type II errors, denoted by C_1 and C_2, respectively. For brevity, let us call "the post-trial cost associated with Type I error" simply the *Type I cost*, and similarly for Type II error, the *Type II cost*.

Asymmetric costs for Type I and Type II errors allow us to incorporate the consequences of these two errors with different weights in our formulation. For example, in the case of a life-threatening disease, when patients can benefit tremendously from an effective drug, the Type II cost—caused by mistakenly rejecting an effective drug—must be much larger than the Type I cost, i.e., $C_1 \ll C_2$. On the other hand, if the disease to be treated is mild and there are several other treatment options available, e.g., mild anemia or secondary infertility, the cost of approving an ineffective drug can be much larger than the cost of not approving an effective drug, i.e., $C_1 \gg C_2$. If the severity of the disease is intermediate, or if no acceptable treatment options exist, e.g., moderate anemia or mild dementia, then these two post-trial costs may be more or less the same, i.e., $C_1 \approx C_2$.

The two post-trial costs, C_1 and C_2, are assumed to be proportional to the size of the target population of the drug. The larger the prevalence of the disease, the higher the cost caused by a wrong decision, and hence the larger the values of C_1 and C_2. If the size of the target population is N, assume there exist two constants, c_1 and c_2, which are independent of the

Table 1
Post-trial and in-trial costs associated with a balanced fixed-sample randomized clinical trial, where $C_1 = Nc_1$ and $C_2 = Nc_2$.

	Post-Trial		In-Trial
	$\widehat{H} = H_0$	$\widehat{H} = H_1$	
$H = H_0$	0	C_1	nc_1
$H = H_1$	C_2	0	$n\gamma C_2$

disease prevalence and depend only on the effectiveness of the drug and the characteristics of the disease, respectively, such that the following linear relation holds:

$$C_i = Nc_i, \quad i = 1, 2, \tag{3}$$

where c_1 and c_2 can be interpreted as the cost per person for Type I and Type II errors, respectively. Lower case letters represent cost per individual, while uppercase letters are used for aggregate costs.

In-trial costs are mainly related to patients' exposure to inferior treatment, e.g., the exposure of enrolled patients to an ineffective drug in the treatment arm or the delay in treating *all* patients (in the control group and in the general population) with an effective drug. If the drug being tested is ineffective, the n subjects in the treatment arm collectively experience an in-trial cost of nc_1. In this case, the patients in the control arm experience no extra cost, since the current treatment or the placebo is assumed to be an acceptable standard of care. However, if the drug is effective, then an unnecessary increase in sample size inevitably leads to a longer trial which implies a delay in getting an effective therapy to patients. This delay affects all patients, both in and outside the trial. Therefore, we model this cost to be a fraction of the aggregate Type II cost, C_2, and linear in the number of subjects in the trial, n. To be more specific, we assign an in-trial cost of $n\gamma C_2$ for an appropriate choice of γ (for the results presented in Section 6, we use $\gamma = 4 \times 10^{-3} \frac{\delta_0}{\sigma}$). All cost categories associated with a fixed-sample test are tabulated in Table 1.

For a given fixed-sample test $\texttt{fxd}(n, \lambda_n)$, where Z_n is observed, and the true underlying hypothesis is H, we can now define the incurred cost, denoted by $C(H, Z_n, \texttt{fxd}(n, \lambda_n))$, as:

$$C(H, Z_n, \texttt{fxd}(n, \lambda_n)) = \begin{cases} Nc_1 \mathbb{1}_{\{Z_n \geq \lambda_n\}} + nc_1, & H = H_0 \\ Nc_2 \mathbb{1}_{\{Z_n < \lambda_n\}} + n\gamma Nc_2, & H = H_1, \end{cases} \tag{4}$$

where $\mathbb{1}$ is the indicator function that takes on the value 1 when its argument is true, and is equal to zero otherwise. Here, the first line corresponds to the case where the drug is ineffective, denoted by $H = H_0$. In this case, there will be a post-trial cost, C_1, due to Type I error, i.e., approving the ineffective drug, which is the first term. The second term in the first line is the in-trial cost of having n patients in the treatment arm taking this ineffective drug. The second line in (4) represents the case in which the drug is effective, denoted by $H = H_1$. In this case, the second term is the in-trial cost, as explained earlier, and the first term is due to rejecting the effective drug, i.e., if $Z_n < \lambda_n$, resulting in the post-trial Type II error cost.

BDA-optimal fixed-sample test

Let us assume prior probabilities of p_0 and p_1 for the null and alternative hypotheses, respectively, i.e., $P(H_0) = p_0$ and $P(H_1) = p_1$, where $p_0, p_1 > 0$ and $p_0 + p_1 = 1$. It is then straightforward to calculate the expected value of the cost, associated with $\texttt{fxd}(n, \lambda_n)$ and given by (4), as the following:

$$\begin{aligned} C(\texttt{fxd}(n, \lambda_n)) &\triangleq E[C(H, Z_n, \texttt{fxd}(n, \lambda_n))] \\ &= p_0 c_1 \left[N\Phi(-\lambda_n) + N\bar{c}_2 \Phi(\lambda_n - \delta_0 \sqrt{\mathcal{I}_n}) + n(1 + \gamma N \bar{c}_2) \right], \end{aligned} \tag{5}$$

where Φ is the cumulative distribution function of a standard normal random variable and E is the expectation operator.

For the remainder of this article, we assume a non-informative prior, i.e., $p_0 = p_1 = 0.5$, in which case $\bar{c}_2 = \frac{p_1 c_2}{p_0 c_1}$ reduces to $\bar{c}_2 = \frac{c_2}{c_1}$, the normalized Type II cost. A non-informative prior is consistent with the "equipoise" principle of two-arm clinical trials (Freedman, 1987): it is only ethical to assign the same number of patients to both arms if there is no prior information in the medical profession that favors one arm over the other. However, in some cases we can formulate more informed priors based on information accumulated through earlier-phase trials and other sources. In such cases, for ethical reasons, the randomization of patients should reflect this information—especially for life-threatening conditions—and the natural framework for doing so is Bayesian adaptive design (Berry, 2006; Barker et al., 2009).

The optimal sample size, n^*, and critical value, λ_n^*, are jointly determined such that the expected cost of the trial, given by (5), is minimized subject to an upper bound for the power level, $\text{Power}_{\max}$, which we set to 90% in our simulations. The power constraint is intended to represent typical practices in the pharmaceutical industry. For example, according to Piantadosi (Piantadosi, 2005, p. 277), "Convention holds that most clinical trials should be designed with a two-sided α-level

set at 0.5 and 80% or 90% power ($\beta = 0.2$ or 0.1, respectively)". Therefore, we have a constrained nonlinear multivariate optimization problem:

$$\left(n^*, \lambda_n^*\right) = \underset{n\in\mathbb{N},\, \lambda_n\in\mathbb{R}}{\operatorname{argmin}}\, C(\texttt{fxd}(n, \lambda_n)) \quad \text{subject to} \quad 1-\beta \leq \text{Power}_{\max}, \tag{6}$$

where $\operatorname{argmin} f$ denotes the minimizer of the function f, and $\mathbb{N}$ and $\mathbb{R}$ denote the set of natural numbers and real numbers, respectively (see Appendix A.1 for a detailed description of the solution to this optimization problem). The fixed-sample test with these two optimal parameters, i.e., $\texttt{fxd}(n^*, \lambda_n^*)$, will be referred to as the BDA-optimal fixed-sample test.

5. Estimating the cost of disease

To estimate the two cost parameters, c_1 and c_2, we use the U.S. Burden of Disease Study 2010 (Murray et al., 2013), which follows the same methodology as that of the comprehensive Global Burden of Disease Study 2010 (GBD 2010), but only using U.S.-level data. We estimate the severity of the adverse effects of medical treatment to define c_1 and the severity of a disease to define c_2.

A key factor in quantifying the burden of disease and loss of health due to disease and injury in the GBD 2010 and the U.S. Burden of Disease Study is the YLD (years lived with disability) attributed to each disease in the study population. To compute YLDs, these studies first specify different sequelae (outcomes) for each specific disease and then multiply the prevalence of each sequela by its disability weight, which is a measure of severity for each sequela and ranges from 0 (no loss of health) to 1 (complete loss of health, i.e., death). For example, the disability weight associated with mild anemia is 0.005; for the terminal phase of cancers without medication, the weight is 0.519. These disability weights are robust across different countries and different social classes (Salomon et al., 2012), and the granularity of the sequelae is such that the final YLD number for the disease is affected by the current status of available treatments for the disease. This makes YLDs especially suitable for our work, since c_2 is the reduction in the severity of the disease to be treated by a new effective therapy, taking into account the current state of available therapies for the disease. We estimate the overall severity of disease using the following relation:

$$s_2 = \frac{D + \text{YLD}}{D + N}, \tag{7}$$

where D is the number of deaths caused by the disease, YLD is the number of YLDs attributed to the disease, and N is the prevalence of the disease in the U.S., all in 2010. It should be noted that YLDs are computed only from non-fatal sequelae. Therefore, to quantify the severity of each disease, we add the number of deaths (multiplied by its disability weight, i.e., 1) to the number of YLDs and divide the result by the number of deaths added to the number of people afflicted with the disease in 2010, hence $D + N$ in the denominator. Furthermore, instead of using the absolute numbers for death, YLD, and prevalence, we use their age-standardized rates (per 100,000) to obtain a severity estimate that is more representative of the severity of the disease in the population. Age standardization is a stratified sampling technique in which different age groups in the population are sampled based on a standard population distribution proposed by the World Health Organization (WHO) (Ahmad et al., 2001). This technique facilitates meaningful comparison of rates for different populations and diseases.

We now assume that the treatment effect $\delta_0 = \sigma$ and that larger treatment effects correspond to a cure, i.e., the cost parameter, c_2, denoting the reduction in the severity of the disease due to an effective drug, is equal to the severity of the disease, s_2. For any treatment effects, δ_0, smaller than σ—because the effective therapy would not completely cure the disease—the cost parameter, c_2, is a fraction of the severity of the disease, s_2, as in the following:

$$c_2 = \min\left(\frac{\delta_0}{\sigma}, 1\right) s_2, \tag{8}$$

where δ_0 is the treatment effect, σ is the standard deviation of the response variables in each arm of the RCT, and s_2 denotes the severity of the disease and is given by (7).

To estimate c_1, which is the current cost of adverse effects of medical treatment per patient, we insert the corresponding values for the adverse effect of medical treatment in the U.S. from the U.S. Burden of Disease Study 2010 (Murray et al., 2013) into (7), and the result is $c_1 = s_1 = 0.07$. The value of c_1 can be made more precise and tailored to the drug candidate being tested if information from earlier clinical phases, e.g., phase 1 and phase 2, is used. However, for simplicity, we only consider a common value for c_1 for all diseases. The parameters used in our proposed model are listed in Table 2, along with their values and the sources on which they are based.

6. BDA-optimal tests for the most deadly diseases

The leading causes of death are determined in Murray et al. (2013) by ranking diseases and injuries based on their associated YLLs (Years of Life Lost due to premature death) in the U.S. in 2010. For our purposes, the following categories, while among the leading causes of premature mortality in the U.S., are omitted, either because they are not diseases, or because they are broad collections (their U.S. YLL ranks are listed in parentheses): road injury (5), self harm (6), interpersonal

Table 2
Parameters used in BDA framework.

Parameter	Value	Description	Source
N	variable	Size of patient population	U.S. Burden of Disease Study (Murray et al., 2013)
c_1	0.07	Cost of side effects per patient	Estimated using the numbers reported in Murray et al. (2013)
c_2	variable	Burden of disease per patient	Estimated using the numbers reported in Murray et al. (2013)
p_0	0.5	Probability that the treatment is ineffective (and possibly toxic)	Assumption (non-informative prior)
δ_0	$2^{-n}\sigma$ $n = 0, 1, 2, 3$	Magnitude of the treatment effect for an effective drug	Assumption (σ is the standard deviation of observations in each arm)
γ	$4 \times 10^{-3}\frac{\delta_0}{\sigma}$	Incremental cost incurred due to adding an extra patient to each arm	Assumption

Table 3
Selected diseases from the 30 leading causes of premature mortality in the U.S., their rank with respect to their U.S. YLL's, prevalence, and severity. The sample size and critical value for the BDA-optimal fixed-sample tests and their size and statistical power at the alternative hypothesis are reported. The alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{8}$. YLL: Number of years of life lost due to premature mortality, BDA: Bayesian decision analysis.

YLL Rank	Disease Name	Prevalence (Thousands)	Severity	Sample Size	Critical Value	α (%)	Power (%)
1	Ischemic heart disease	8,895.61	0.12	806	1.853	3.2	74.4
2	Lung cancer	289.87	0.45	521	1.094	13.7	82.2
3a	Ischemic stroke	3,932.33	0.15	767	1.755	4.0	75.6
3b	Hemorrhagic/other non-ischemic stroke	949.33	0.16	751	1.723	4.2	75.8
4	Chronic obstructive pulmonary disease	32,372.11	0.06	940	2.182	1.5	70.1
7	Diabetes	23,694.90	0.05	958	2.226	1.3	69.5
8	Cirrhosis of the liver	78.37	0.49	491	1.024	15.3	82.5
9	Alzheimer's disease	5,145.03	0.18	727	1.652	4.9	76.8
10	Colorectal cancer	798.90	0.15	752	1.727	4.2	75.7
11a	Pneumococcal pneumonia	84.14	0.30	596	1.351	8.8	79.0
11b	Influenza	119.03	0.20	679	1.584	5.7	76.4
11c	H influenzae type B pneumonia	21.15	0.26	545	1.378	8.4	75.4
11d	Respiratory syncytial virus pneumonia	14.90	0.07	–	–	–	–
13	Breast cancer	3,885.25	0.05	951	2.218	1.3	69.4
16	Chronic kidney disease	9,919.02	0.04	981	2.288	1.1	68.5
18	Pancreatic cancer	22.67	0.71	384	0.711	23.9	84.6
20	Cardiomyopathy	416.31	0.17	729	1.677	4.7	76.1
21	Hypertensive heart disease	185.26	0.27	633	1.429	7.6	78.7
22	Leukemia	139.75	0.21	671	1.551	6.0	77.0
23	HIV/AIDS	1,159.58	0.10	830	1.926	2.7	73.3
24	Kidney cancers	328.94	0.12	794	1.864	3.1	73.4
25	Non-Hodgkin lymphoma	282.94	0.13	766	1.792	3.7	74.4
27	Prostate cancer	3,709.70	0.05	967	2.259	1.2	68.8
28	Brain and nervous system cancers	59.76	0.30	585	1.339	9.0	78.8
30	Liver cancer	31.27	0.44	492	1.080	14.0	81.1

violence (12), pre-term birth complications (14), drug-use disorders (15), other cardiovascular/circulatory diseases (17), congenital anomalies (19), poisonings (26), and falls (29). We also divided two categories into subcategories: stroke is split into ischemic stroke and non-ischemic stroke, and lower respiratory tract infections are divided into four diseases (11a–d in Tables 3–6). These choices yield 25 leading causes of death for which we compute BDA-optimal thresholds.

The estimated severity for each disease, s_2, is reported in the fourth column of Table 3. As can be seen, some cancers are not quite as severe as other non-cancerous diseases. For instance, prostate cancer ($s_2 = 0.05$), is much less harmful than cirrhosis ($s_2 = 0.49$), which must be due to the current state of medication for prostate cancer and the lack of any effective treatment for cirrhosis in the U.S. On the other hand, some cancers are shown to be extremely deadly, e.g., pancreatic cancer, with $s_2 = 0.71$. Using this measure of severity, we have an objective data-driven framework where different diseases with different afflicted populations can be compared to one another.

Table 4
Selected diseases from the 30 leading causes of premature mortality in the U.S., their rank with respect to their U.S. YLL's, prevalence, and severity. The sample size and critical value for the BDA-optimal fixed-sample tests and their size and statistical power at the alternative hypothesis are reported. The alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{4}$. YLL: Number of years of life lost due to premature mortality, BDA: Bayesian decision analysis.

YLL Rank	Disease Name	Prevalence (Thousands)	Severity	Sample Size	Critical Value	α (%)	Power (%)
1	Ischemic heart disease	8,895.61	0.12	283	1.759	3.9	88.8
2	Lung cancer	289.87	0.45	165	0.989	16.1	90.0
3a	Ischemic stroke	3,932.33	0.15	268	1.656	4.9	89.2
3b	Hemorrhagic/other non-ischemic stroke	949.33	0.16	263	1.624	5.2	89.3
4	Chronic obstructive pulmonary disease	32,372.11	0.06	336	2.098	1.8	87.3
7	Diabetes	23,694.90	0.05	344	2.144	1.6	87.2
8	Cirrhosis of the liver	78.37	0.49	156	0.926	17.7	90.0
9	Alzheimer's disease	5,145.03	0.18	253	1.550	6.1	89.7
10	Colorectal cancer	798.90	0.15	264	1.630	5.2	89.3
11a	Pneumococcal pneumonia	84.14	0.30	205	1.250	10.6	90.0
11b	Influenza	119.03	0.20	243	1.488	6.8	89.8
11c	H influenzae type B pneumonia	21.15	0.26	214	1.304	9.6	90.0
11d	Respiratory syncytial virus pneumonia	14.90	0.07	278	1.928	2.7	84.6
13	Breast cancer	3,885.25	0.05	342	2.135	1.6	87.2
16	Chronic kidney disease	9,919.02	0.04	354	2.207	1.4	86.9
18	Pancreatic cancer	22.67	0.71	117	0.631	26.4	90.0
20	Cardiomyopathy	416.31	0.17	257	1.579	5.7	89.5
21	Hypertensive heart disease	185.26	0.27	218	1.329	9.2	90.0
22	Leukemia	139.75	0.21	239	1.454	7.3	90.0
23	HIV/AIDS	1,159.58	0.10	295	1.837	3.3	88.5
24	Kidney cancers	328.94	0.12	285	1.777	3.8	88.6
25	Non-Hodgkin lymphoma	282.94	0.13	274	1.701	4.4	89.0
27	Prostate cancer	3,709.70	0.05	349	2.177	1.5	87.0
28	Brain and nervous system cancers	59.76	0.30	204	1.243	10.7	90.0
30	Liver cancer	31.27	0.44	165	0.989	16.1	90.0

Table 5
Selected diseases from the 30 leading causes of premature mortality in the U.S., their rank with respect to their U.S. YLL's, prevalence, and severity. The sample size and critical value for the BDA-optimal fixed-sample tests and their size and statistical power at the alternative hypothesis are reported. The alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{2}$. YLL: Number of years of life lost due to premature mortality, BDA: Bayesian decision analysis.

YLL Rank	Disease Name	Prevalence (Thousands)	Severity	Sample Size	Critical Value	α (%)	Power (%)
1	Ischemic heart disease	8,895.61	0.12	73	1.739	4.1	90.0
2	Lung cancer	289.87	0.45	41	0.982	16.3	90.0
3a	Ischemic stroke	3,932.33	0.15	69	1.655	4.9	90.0
3b	Hemorrhagic/other non-ischemic stroke	949.33	0.16	67	1.612	5.3	90.0
4	Chronic obstructive pulmonary disease	32,372.11	0.06	90	2.073	1.9	90.0
7	Diabetes	23,694.90	0.05	93	2.128	1.7	90.0
8	Cirrhosis of the liver	78.37	0.49	39	0.926	17.7	90.0
9	Alzheimer's disease	5,145.03	0.18	64	1.547	6.1	90.0
10	Colorectal cancer	798.90	0.15	68	1.634	5.1	90.0
11a	Pneumococcal pneumonia	84.14	0.30	52	1.268	10.2	90.0
11b	Influenza	119.03	0.20	61	1.480	6.9	90.0
11c	H influenzae type B pneumonia	21.15	0.26	54	1.317	9.4	90.0
11d	Respiratory syncytial virus pneumonia	14.90	0.07	84	1.959	2.5	90.0
13	Breast cancer	3,885.25	0.05	92	2.110	1.7	90.0
16	Chronic kidney disease	9,919.02	0.04	96	2.183	1.5	90.0
18	Pancreatic cancer	22.67	0.71	30	0.655	25.6	90.0
20	Cardiomyopathy	416.31	0.17	65	1.569	5.8	90.0
21	Hypertensive heart disease	185.26	0.27	54	1.317	9.4	90.0
22	Leukemia	139.75	0.21	60	1.457	7.3	90.0
23	HIV/AIDS	1,159.58	0.10	77	1.821	3.4	90.0
24	Kidney cancers	328.94	0.12	74	1.760	3.9	90.0
25	Non-Hodgkin lymphoma	282.94	0.13	71	1.698	4.5	90.0
27	Prostate cancer	3,709.70	0.05	95	2.164	1.5	90.0
28	Brain and nervous system cancers	59.76	0.30	51	1.243	10.7	90.0
30	Liver cancer	31.27	0.44	42	1.010	15.6	90.0

Table 6
Selected diseases from the 30 leading causes of premature mortality in the U.S., their rank with respect to their U.S. YLL's, prevalence, and severity. The sample size and critical value for the BDA-optimal fixed-sample tests and their size and statistical power at the alternative hypothesis are reported. The alternative hypothesis corresponds to $\delta_0 = \sigma$. YLL: Number of years of life lost due to premature mortality, BDA: Bayesian decision analysis.

YLL Rank	Disease Name	Prevalence (Thousands)	Severity	Sample Size	Critical Value	α (%)	Power (%)
1	Ischemic heart disease	8,895.61	0.12	18	1.718	4.3	90.0
2	Lung cancer	289.87	0.45	10	0.955	17.0	90.0
3a	Ischemic stroke	3,932.33	0.15	17	1.634	5.1	90.0
3b	Hemorrhagic/other non-ischemic stroke	949.33	0.16	17	1.634	5.1	90.0
4	Chronic obstructive pulmonary disease	32,372.11	0.06	23	2.110	1.7	90.0
7	Diabetes	23,694.90	0.05	23	2.110	1.7	90.0
8	Cirrhosis of the liver	78.37	0.49	10	0.955	17.0	90.0
9	Alzheimer's disease	5,145.03	0.18	16	1.547	6.1	90.0
10	Colorectal cancer	798.90	0.15	17	1.634	5.1	90.0
11a	Pneumococcal pneumonia	84.14	0.30	13	1.268	10.2	90.0
11b	Influenza	119.03	0.20	15	1.457	7.3	90.0
11c	H influenzae type B pneumonia	21.15	0.26	14	1.364	8.6	90.0
11d	Respiratory syncytial virus pneumonia	14.90	0.07	21	1.959	2.5	90.0
13	Breast cancer	3,885.25	0.05	23	2.110	1.7	90.0
16	Chronic kidney disease	9,919.02	0.04	24	2.183	1.5	90.0
18	Pancreatic cancer	22.67	0.71	7	0.589	27.8	90.0
20	Cardiomyopathy	416.31	0.17	16	1.547	6.1	90.0
21	Hypertensive heart disease	185.26	0.27	14	1.364	8.6	90.0
22	Leukemia	139.75	0.21	15	1.457	7.3	90.0
23	HIV/AIDS	1,159.58	0.10	19	1.801	3.6	90.0
24	Kidney cancers	328.94	0.12	19	1.801	3.6	90.0
25	Non-Hodgkin lymphoma	282.94	0.13	18	1.718	4.3	90.0
27	Prostate cancer	3,709.70	0.05	24	2.183	1.5	90.0
28	Brain and nervous system cancers	59.76	0.30	13	1.268	10.2	90.0
30	Liver cancer	31.27	0.44	10	0.955	17.0	90.0

Having estimated the severity of different diseases, we apply the methodology introduced in Section 4 to determine BDA-optimal fixed-sample tests for testing drugs intended to treat each disease. The sample size, critical value, size, and statistical power of these BDA-optimal tests are reported in Tables 3–6 for cases where the alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{8}$, $\delta_0 = \frac{\sigma}{4}$, $\delta_0 = \frac{\sigma}{2}$, and $\delta_0 = \sigma$, respectively.

Some of the diseases listed in Tables 3–6 are no longer considered a single disease, but rather a collection of diseases with heterogeneous biological and genetic profiles, and with distinct patient populations (Polyak, 2011; Berry, 2015b), e.g., breast cancer. This trend towards finer and finer stratifications is particularly relevant for oncology, where biomarkers have subdivided certain types of cancer into many subtle but important variations (Berry, 2015b). However, because data on the burden of disease are not yet available for these subdivisions, we use the conventional categories in Tables 3–6, i.e., where each cancer type is decided based on the site of the tumor.

The reported values for the power of BDA-optimal tests increase with the severity of the disease and its prevalence. This is because the overall burden of disease ($C_2 = Nc_2$) is large either when the cost parameter, c_2, is large due to the large severity of the disease, s_2, e.g., pancreatic cancer, or when the disease is highly prevalent (large N), e.g., prostate cancer. This is true for life-threatening orphan diseases that have small populations ($N < 200{,}000$ in the U.S.) but large severity (s_2). This has recently become noteworthy because many cancers are being reclassified as orphan diseases through the use of biomarkers and personalized medicine (Berry, 2015b). By this measure, not approving an effective drug is a costly option, hence BDA-optimal tests exhibit higher power for diseases with a larger overall burden of disease to detect positive treatment effects.

The general dependence of the statistical power on the overall burden of disease, i.e., its prevalence multiplied by its severity, can be observed in Fig. 1. The shaded area in Fig. 1(a) is the set of prevalence/severity pairs for which *not* conducting an RCT is better than running any fixed-sample balanced two-arm RCTs. For these diseases, the minimum of the cost function in (5) has a higher cost than forgoing the benefits of a potential therapy whose treatment effect is $\delta_0 = \frac{\sigma}{8}$. In other words, for these diseases, the combination of the small treatment effect assumed for the effective therapy, the mild severity of the disease, and its intermediate prevalence cannot justify running the type of RCT that we consider in this article. This is expected, however, because small treatment effects may not be clinically significant for mild diseases (McGlothlin and Lewis, 2014). As seen in panel (a) of Figs. 4–6, where we assume the alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{4}$, this "no-RCT" area shrinks as the treatment effect increases, since the benefit that patients would gain in the case of an effective drug is higher than when the treatment effect is $\delta_0 = \frac{\sigma}{8}$, *ceteris paribus*.

In Fig. 1(a), the contour plot of the power of BDA-optimal tests is presented, where most of the contour lines coincide with constant overall burdens of disease, i.e., $Ns_2 =$ constant, which are straight lines with negative slope on a log–log graph. Also, to facilitate visualizing where each disease in Table 3 lies in the prevalence–severity plane, we have superimposed the

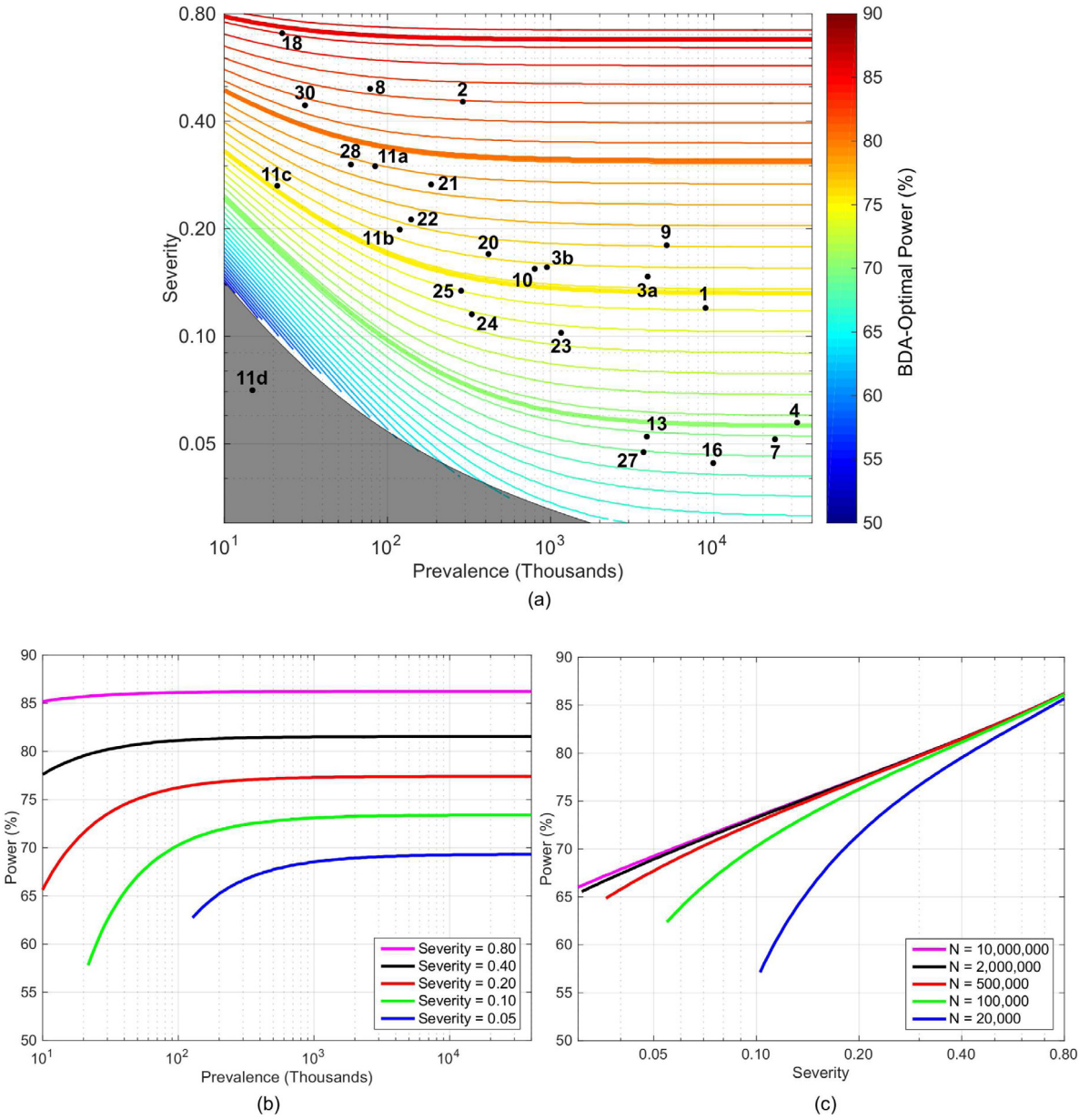

Fig. 1. The statistical power of the BDA-optimal fixed-sample test at the alternative hypothesis corresponding to $\delta_0 = \frac{\sigma}{8}$. Panel (a) shows the contour levels for the power, while panels (b) and (c) demonstrate its cross-sections along the two axes. The contour lines corresponding to the power levels $1 - \beta = 70\%, 75\%, 80\%$, and 85% are highlighted in panel (a). The superimposed numbers in panel (a) denote the YLL rank of each disease in Table 3. YLL: Number of years of life lost due to premature mortality, BDA: Bayesian decision analysis.

YLL rank of each disease in Fig. 1(a). For example, pancreatic cancer is number 18, which has the highest severity among the listed diseases. We have also included the cross-sections of power for BDA-optimal tests in Fig. 1(b) and 1(c).

In sharp contrast to the narrow range of power for the BDA-optimal tests in Table 3, the size of these tests varies dramatically across different diseases. As seen in Table 3, with few exceptions, the size of the test mainly depends on the severity of the disease. In general, as the severity of the disease increases, the critical value to approve the drug becomes less conservative, i.e., it becomes smaller. This is because the cost per patient of not approving an effective drug becomes much larger than the cost per patient associated with adverse side effects. Consequently, the probability of Type I error, i.e., the size of the test, increases. For example, for pancreatic cancer, the critical value is as low as 0.711, while for the conventional 2.5%-level fixed-sample test, it is 1.960. This results in a relatively high size (23.9%) for the BDA-optimal test for a drug intended

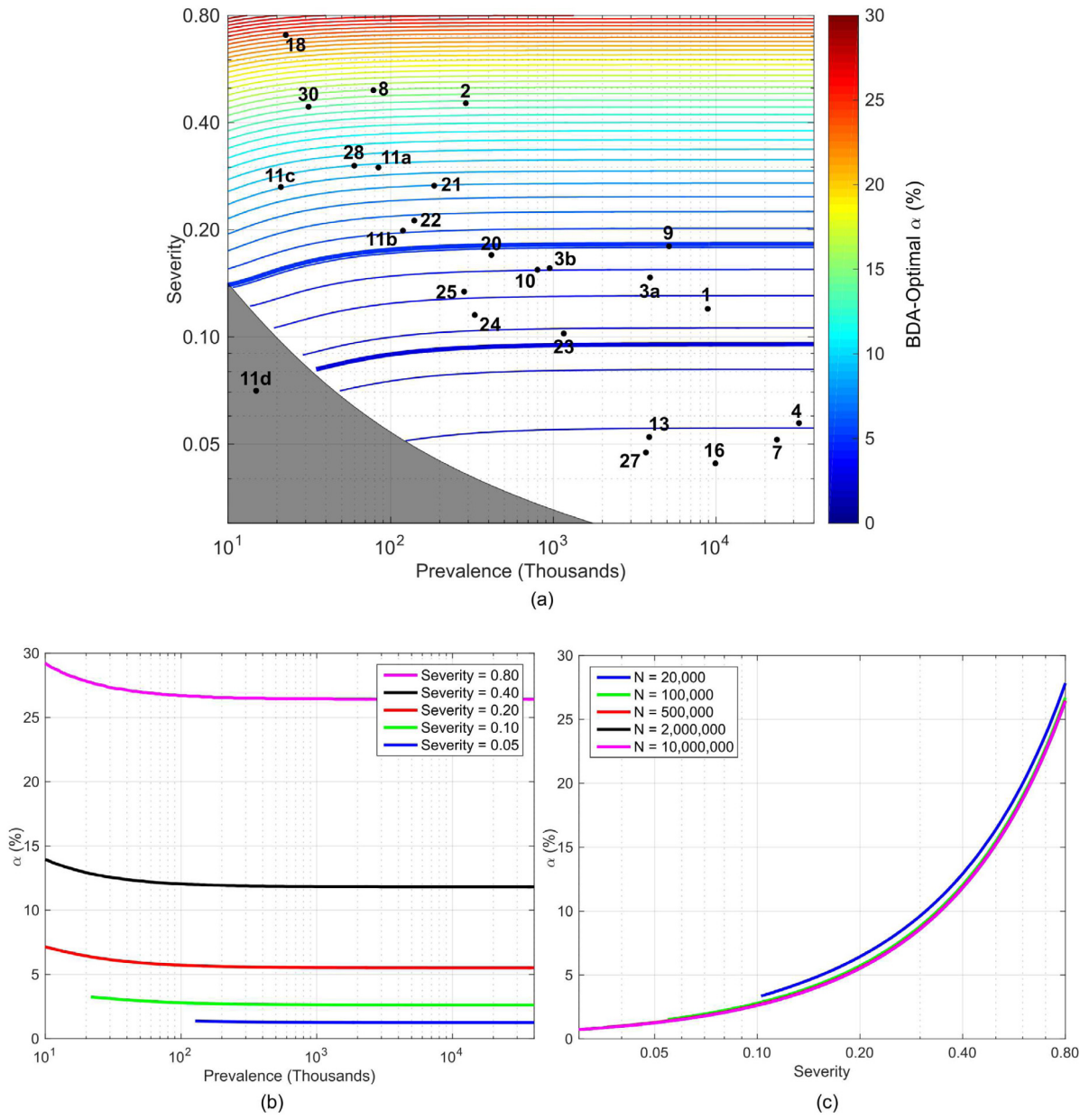

Fig. 2. The size of the BDA-optimal fixed-sample test as a function of disease severity and prevalence where the alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{8}$. Panel (a) shows the contour levels for the size, while panels (b) and (c) demonstrate its cross-sections along the two axes. The contour lines corresponding to $\alpha = 2.5\%$ and $\alpha = 5.0\%$ are highlighted in panel (a). The superimposed numbers in panel (a) denote the YLL rank of each disease in Table 3. YLL: Number of years of life lost due to premature mortality, BDA: Bayesian decision analysis.

to treat pancreatic cancer, consistent with the urgency of approving drugs intended to treat life-threatening diseases that have no existing effective treatment.

However, it should be noted that the conventional value of 2.5% for the probability of Type I error, while too conservative for terminal diseases, is not conservative enough for less severe diseases, e.g., diabetes, for which the size of the BDA-optimal test is 1.3%. The size of BDA-optimal tests for a large range of severity and prevalence values is presented in Fig. 2. The size monotonically increases with disease severity for any given prevalence, and as seen in Fig. 2(a) and 2(b), it becomes independent of the prevalence for all target populations with more than 200,000 patients, hence the horizontal contour lines for x values larger than 200 in Fig. 2(a). This insensitivity of the size to disease prevalence makes our model quite robust to estimation error in this parameter.

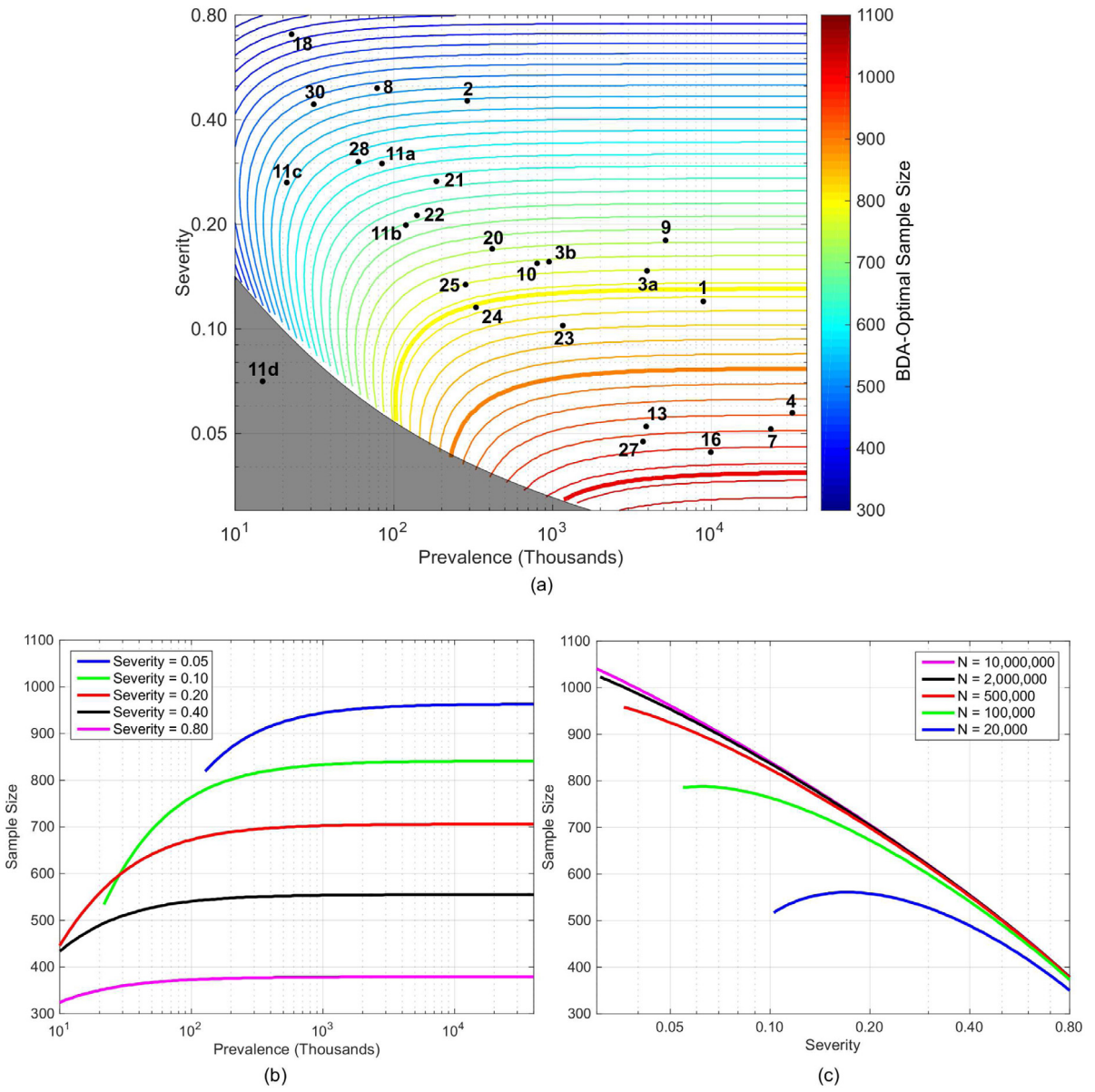

Fig. 3. The sample size of the BDA-optimal fixed-sample test for different severity and prevalence values where the alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{8}$. Panel (a) shows the contour levels for the sample size, while panels (b) and (c) demonstrate its cross-sections along the two axes. The contour lines associated with the sample size of conventional fixed-sample tests with $\alpha = 2.5\%$ and $1 - \beta = 70\%$, 75%, 80%, and 85% are highlighted in panel (a). The superimposed numbers in panel (a) denote the YLL rank of each disease in Table 3. YLL: Number of years of life lost due to premature mortality, BDA: Bayesian decision analysis.

It is useful to investigate the dependence of the sample size of BDA-optimal tests on the prevalence and severity of disease. First, we observe in Fig. 3(b) that, for any given severity value, the sample size of the BDA-optimal test increases with the prevalence of the disease. This supports the intuitive argument that the sample size should increase with the size of the target population. Furthermore, a unique trend is observed in Fig. 3(c): As the severity of the disease increases, for a large enough target population ($N > 100{,}000$), the optimal sample size continuously shrinks to avoid any delay in bringing the effective drug into the market because of the high toll ($C_2 = Nc_2$) that the disease has on society.

On the other hand, for relatively small populations, e.g., $N = 20{,}000$, the optimal sample size peaks somewhere in the middle of the severity spectrum. This occurs because of two opposing trends. The disease burden on society is quite low for small populations and a disease of low severity; therefore, exposure to a potentially toxic treatment in the trial is not worth the risk. Under these conditions, the sample size should be as small as possible. However, for small populations and a disease of high severity, i.e., a large overall burden of disease, the risk of taking an inferior treatment in the trial becomes

much smaller than that of waiting for an effective treatment to be approved. Hence, the sample size for $N = 20{,}000$ over very large severity values decreases as severity increases. This has particular relevance for rare but severely debilitating or mortal diseases such as amyotrophic lateral sclerosis, Duchenne muscular dystrophy, or Gaucher disease.

In between these two extremes, where the overall burden of disease is not that high, and the disease has intermediate severity, the sample size of the trial is allowed to become larger to guarantee an appropriate balance between approving an effective drug as fast as possible and not exposing the patients to a drug with adverse side effects.

It is worth emphasizing that, as with the size of the test, the sample size of BDA-optimal tests is quite insensitive to disease prevalence for large target populations (hence the horizontal contour lines in Fig. 3(a) over large values of prevalence), which suggests that these results are robust. In particular, in Figs. 4–6, we present the corresponding BDA-optimal power, size, and sample size, respectively, for the alternative hypothesis $\delta_0 = \frac{\sigma}{4}$. The same general trends as in Figs. 1–3 are observed.

7. Qualifications and extensions

A number of outstanding issues must still be addressed before BDA can be implemented. The most pressing issue is the measurement of the costs and benefits used to establish the tradeoff between Type I and Type II errors. Our use of burden of disease data is merely a starting point; more refined measures can be constructed with additional data that are specific to a given trial. For example, we have not taken into account the duration of a clinical trial or the difficulty in recruiting patients. Both of these considerations impose costs on patients inside and outside of a clinical trial and should be reflected in the cost parameters. In a separate study, we show how such cost parameters can be calibrated in the specific context of ten current clinical trials in oncology (Montazerhodjat et al., 2017).

Incorporating patients as a stakeholder group in determining the value of costs and benefits will be vital to their accurate and transparent measurement. The 2012 Food and Drug Administration Safety and Innovation Act (FDASIA) (U.S. Congress, 2012) has "recognized the value of patient input to the entire drug development enterprise, including FDA review and decision-making". Moreover, Section 3002 of the recently passed 21st Century Cures Act requires the FDA to develop guidelines for patient-focused drug development, which includes collecting patient preference and experience data and incorporating this information in the drug approval process.

One proposal for implementing these mandates is for the FDA to create a patient advisory board consisting of representatives from patient advocacy groups, with the specific charge of formulating explicit cost estimates of Type I and Type II errors. These estimates can then be incorporated into the FDA decision-making process, not mechanically, but as additional inputs into the FDA's quantitative and qualitative deliberations.

To incorporate other perspectives from the entire biomedical ecosystem, the membership of this advisory board could be expanded to include representatives from other stakeholder groups—caregivers, physicians, biopharma executives, regulators, and policymakers. With an expanded composition, this advisory board could play an even broader role than the concept of a Citizens Council adopted by NICE.[6] The diverse set of stakeholders can provide crucial input to the FDA and EMA, reflecting the overall view of society on critical cost parameters. However, the role of such a committee should then be limited to advice; drug approval decisions should be made solely by FDA officials. The separation of recommendations and final decisions helps ensure that the adaptive nature of the proposed framework will not be exploited or gamed by any one party. Moreover, such a framework can provide drug developers with greater certainty about the process by which regulatory agencies incorporate patient values. By adopting this same framework, drug developers can design more efficient clinical trials.

The BDA framework also fills a need mandated by the fifth authorization of the Prescription Drug User Fee Act (PDUFA) for an enhanced quantitative approach to the benefit–risk assessment of new drugs (U.S. Food and Drug Administration, 2013a). Due to its quantitative nature, BDA provides transparency, consistency, and repeatability to the review process, which is one of the key objectives in PDUFA. The sensitivity of the final judgment to the underlying assumptions, e.g., cost vs. benefit, can be evaluated and made available to the public, which renders the proposed framework even more transparent. However, the ability to incorporate prior information and qualitative judgments about relative costs and benefits within the BDA framework preserves important flexibility for regulatory decision-makers.

Prior information is a key input in the BDA framework, one which we have ignored by employing an uninformed prior. However, because of its role as the trusted intermediary in evaluating and approving drug applications, the FDA is privy to information about current industry activity and technology that no other party possesses. As a result, the FDA is in the unique position of being able to formulate highly informed priors on various therapeutic targets, mechanisms, and R&D agendas. Applying such priors in the BDA framework could yield very different outcomes from the uniform priors we used in Section 6, which assumes a 50/50 chance that a drug candidate is effective. While 50/50 may seem more equitable, from a social welfare perspective it is highly inefficient, potentially allowing many more expensive clinical trials to be conducted than necessary.

Although the FDA cannot be expected to play the role of social planner, and while it should be industry-neutral in its review process, ignoring scientific information in favor of 50/50 does not necessarily serve any stakeholder's interests. Moreover, the use of 50/50 when more informative priors are available could be considered unethical in cases involving therapies for terminal illnesses. For example, for pancreatic cancer, if the prior probability of efficacy is 60% instead of 50%,

[6] See https://www.nice.org.uk/Get-Involved/Citizens-Council.

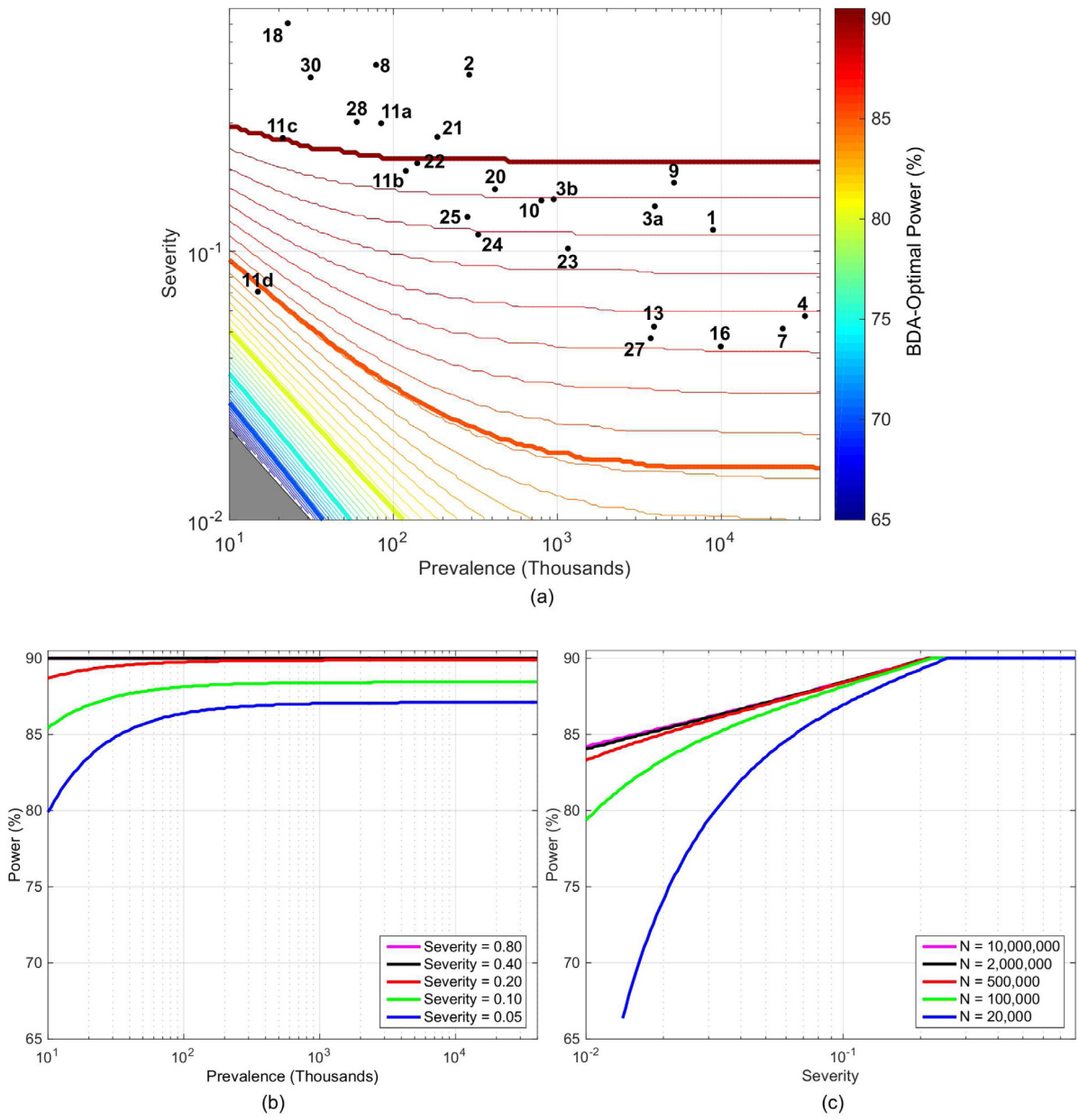

Fig. 4. The statistical power of the BDA-optimal fixed-sample test at the alternative hypothesis corresponding to $\delta_0 = \frac{\sigma}{4}$. Panel (a) shows the contour levels for the power, while panels (b) and (c) demonstrate its cross-sections along the two axes. The contour lines corresponding to the power levels $1 - \beta = 70\%, 75\%, 80\%, 85\%$, and 90% are highlighted in panel (a). The superimposed numbers in panel (a) denote the YLL rank of each disease in Table 4. YLL: Number of years of life lost due to premature mortality, BDA: Bayesian decision analysis.

the size of the BDA-optimal test would be 39.3% rather than 23.9%, leading to many more approvals of therapies that treat the disease. Because of the sensitivity of the results to parameters in the model, it is crucial to use as much available information as possible to arrive at reliable estimates for the costs and other parameter values. Bayesian adaptive designs can play a significant role in ensuring the proper update of available knowledge as more observations—both in and outside the ongoing RCT—are collected and analyzed.

Although the Bayesian adaptive framework is beyond the scope of our current analysis, BDA can easily be extended to adaptive designs. For Bayesian trials, our framework leads to a two-arm bandit problem where the cost function proposed here does not conform to the commonly used cost/reward functions (Berry and Fristedt, 1985). In these trials, at every interim look at the data, the investigator is faced with a decision: to sample or not to sample? If further sampling is required, the

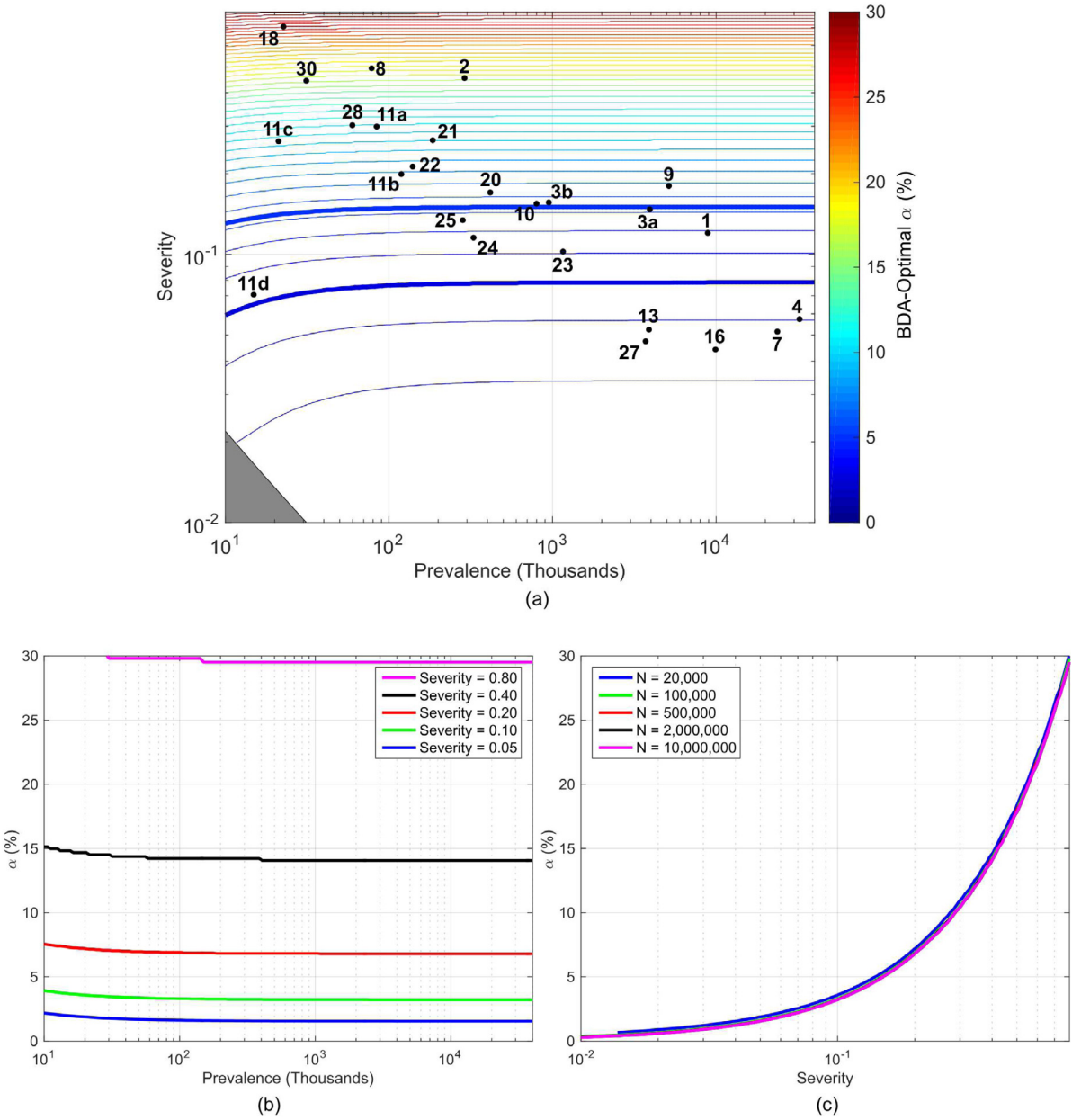

Fig. 5. The size of the BDA-optimal fixed-sample test as a function of disease severity and prevalence where the alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{4}$. Panel (a) shows the contour levels for the size, while panels (b) and (c) demonstrate its cross-sections along the two axes. The contour lines corresponding to $\alpha = 2.5\%$ and $\alpha = 5.0\%$ are highlighted in panel (a). The superimposed numbers in panel (a) denote the YLL rank of each disease in Table 4. YLL: Number of years of life lost due to premature mortality, BDA: Bayesian decision analysis.

investigator must then determine how many additional samples should be collected and what the threshold should be to reject the null hypothesis, with the constraint that the cost to current and future patients should be minimized.

This adaptive process results in a path-dependent optimal strategy in which further sampling may or may not be needed, depending on the promise shown by current observations. At every step, there will be a current cost to making a decision based solely on the currently available data, and an "optimal" expected cost of further sampling. If the optimal expected cost of additional sampling is smaller than the cost to be incurred by making a decision without further sampling, then the investigator continues enrolling patients into the RCT. Otherwise, the trial is stopped, with a decision made based on the currently available data. The technicalities of Bayesian posterior updates and the dynamic programming nature of the

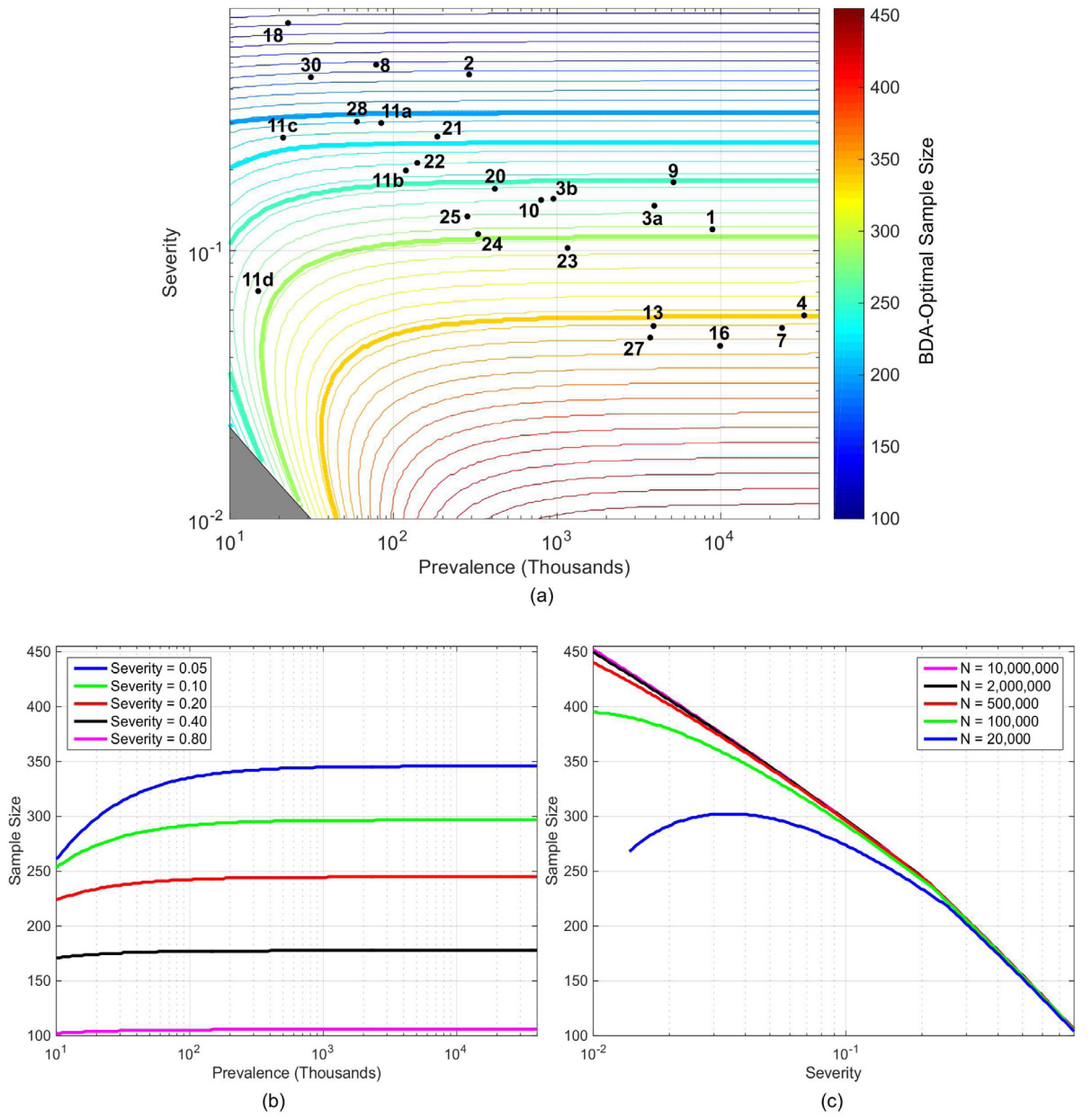

Fig. 6. The sample size of the BDA-optimal fixed-sample test for different severity and prevalence values where the alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{4}$. Panel (a) shows the contour levels for the sample size, while panels (b) and (c) demonstrate its cross-sections along the two axes. The contour lines associated with the sample size of conventional fixed-sample tests with $\alpha = 2.5\%$ and $1 - \beta = 70\%$, 75%, 80%, 85%, and 90% are highlighted in panel (a). The superimposed numbers in panel (a) denote the YLL rank of each disease in Table 4. YLL: Number of years of life lost due to premature mortality, BDA: Bayesian decision analysis.

problem with Bayesian adaptive trials mask the simplicity and importance of the results presented here, however, and we leave the formalization of this approach to future research.

Drug approval is not always a binary choice. The variability of drug response in patient populations—attributed to biological and behavioral factors—has been recognized as a critical element in causing uncertainty and creating the so-called "efficacy-effectiveness" gap (Eichler et al., 2011), where efficacy refers to therapeutic performance in a clinical trial and effectiveness refers to performance in practice. Several proposals have been made for integrated clinical trial pathways to bridge this gap (Selker et al., 2014). In these cases, the BDA framework can be extended by defining costs for a more fine-grained set of events.

Moreover, new paradigms have been proposed to further address the risk associated with the current approval process, e.g., staggered approval (Eichler et al., 2008; European Medicines Agency, 2010) and adaptive licensing (Eichler et al., 2012), which the EMA is actively pursuing (European Medicines Agency, 2014). Adaptive pathways have great potential to benefit all key stakeholders (Baird et al., 2013). In fact, one design principle called for by Eichler et al. (2012) is the employment of less stringent statistical significance levels in efficacy trials for drugs that target life-threatening diseases or rare conditions. Our BDA framework provides an explicit quantitative method for implementing this principle, further motivation for employing it in the drug approval process.

The issue of less stringent approval standards is an important one to consider within the BDA framework. Assuming that more compounds would be developed to treat life-threatening diseases due to the increased incentives for development, the prior probability of efficacy for a sample compound would naturally decrease, since more compounds would be claimed to result in a clinically meaningful effect, increasing the denominator. This decrease in the prior probability of efficacy, or equivalently, the increase in the prior probability of futility, p_0, would require more conservative and less permissive regulatory criteria, ultimately counterbalancing the impact of the high disease severity on the regulatory standards. In the steady state of the new equilibrium of our proposed framework, the "optimal" approval thresholds for life-threatening diseases would be more conservative than the numbers reported in this article. During the transition, a practical approach to protect the public against excess side effects or ineffective treatments would be to adopt newly proposed multi-stage approval paradigms (Eichler et al., 2012), where the treatment for a life-threatening disease is granted a conditional approval. Only after showing consistent efficacy and lack of serious side effects in a larger population would the drug be granted full approval. Otherwise, the drug will be taken off the market.

A closely related issue is the fact that the laxer BDA approval thresholds will, by design, yield more false positives for all treatments (not merely life-threatening ones), and therefore will have the potential for a greater number of patients with adverse side effects. This can be addressed by conducting more vigilant post-approval surveillance, and imposing greater post-approval commitments and requirements for drug and device companies to provide the FDA with data on patient experiences following approval. Failure to provide this data, or mounting evidence of ineffectiveness or severe side effects, would be grounds for revoking the approval. Because experience has shown that revoking an approved drug can be extremely challenging, however, any implementation of BDA should be coupled with the creation of a new category of temporary FDA approval for "Speculative Therapies" (Lo, 2017).

A Speculative Therapies program would be a middle ground between a clinical trial and an approval, similar in spirit to expanded access programs[7] and the adaptive designs of sophisticated clinical trials with master protocols such as I-SPY 2 (Harrington and Parmagiani, 2016), LUNG-MAP (Steuer et al., 2015), and GBM-AGILE (National Biomarker Development Alliance, 2017), in which patient care and clinical investigations are simultaneously accomplished.

This new category would offer a limited two-year license to market a speculative therapy to a pre-specified patient population, mandating no off-label use of the therapy, with regular monitoring and data reports sent to the FDA by the manufacturer and/or patients' physicians during this period. At the end of the two-year period, one of three outcomes would occur: (a) the manufacturer would be able to apply for a second two-year license (only one renewal will be allowed); (b) the license would expire; or (c) the therapy would receive the traditional FDA approval designation. Of course, at any point during the two-year period, the FDA would be able to terminate the license if the accumulated data suggests an ineffective or unsafe therapy. While this process may impose greater burdens on patients, manufacturers, and the FDA, it may be worthwhile if it brings quicker relief to patients facing mortal illnesses and extreme suffering.

Finally, reducing the regulatory hurdle for treatments targeting life-threatening diseases with no available therapies should encourage pharmaceutical companies to shift their focus and investments toward life-threatening diseases. However, this might have the unintended consequence that the companies would compensate for this additional investment in those areas by reducing their R&D dollars in other, less serious, disease areas. This is a valid concern, but even under the current drug approval paradigm, less severe conditions are often neglected by the industry because of disincentives related to insurance coverage and reimbursement policies, rather than regulation. Incorporating patient preferences into the drug approval process may help pharma focus on developing therapies that are most meaningful to patients.

8. Conclusion

To address the inflexibility of traditional frequentist designs for randomized clinical trials, we propose an optimal fixed-sample test within a BDA framework that incorporates both the potential asymmetry in the costs of Type I and Type II errors, and the costs of ineffective treatment during and after the trial. Our findings suggest that conventional standards of statistical significance for approving drugs may be overly conservative for the most deadly diseases and overly aggressive for the mildest ones. Therefore, changing the "one size fits all" statistical criteria for drug approval is likely to yield greater benefits to a greater portion of the population, as demonstrated by others (Deley et al., 2012).

The FDA and many of its foreign counterparts already take into account a variety of factors beyond p-values in making their decisions. The recent controversial FDA approval of the Duchenne muscular dystrophy drug eteplirsen (Exondys 51), despite relatively weak clinical evidence, suggests that the FDA does consider the patient's perspective in their process. However, their deliberations are largely opaque—even to industry insiders—and the exact role and weight of patient preferences are unclear. BDA provides an explicit, systematic, objective, transparent, and repeatable framework for explicitly incorporating such preferences, as well as data on the burden of disease, into the therapeutic approval process.

[7] See, for example, the FDA's "compassionate use" program: https://www.fda.gov/NewsEvents/PublicHealthFocus/ExpandedAccessCompassionateUse/default.htm.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2018.12.009.

References

Ahmad, O.B., Boschi-Pinto, C., Lopez, A.D., et al., 2001. Age standardization of rates: A new WHO standard. GPE discussion paper series: No 31. http://www.who.int/healthinfo/paper31.pdf. (Accessed 20 July 2014).

American Cancer Society. Jan. 2015. Pancreatic cancer survival by stage. http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-survival-rates. (Accessed 23 January 2016).

Anscombe, F.J., 1963. Sequential medical trials. J. Am. Stat. Assoc. 58 (302), 365–383.

Aronson, J.K., 2008. Drug withdrawals because of adverse effects. In: Aronson, J.K. (Ed.), A Worldwide Yearly Survey of New Data and Trends in Adverse Drug Reactions and Interactions. In: Side Effects of Drugs Annual, vol. 30, Elsevier, http://dx.doi.org/10.1016/S0378-6080(08)00064-0, pp. xxxi–xxxv.

Baird, L.G., Trusheim, M.R., Eichler, H.G., Berndt, E.R., Hirsch, G., 2013. Comparison of stakeholder metrics for traditional and adaptive development and licensing approaches to drug development. Ther. Innov. Regul. Sci 47 (4), 474–483.

Barker, A.D., Sigman, C.C., Kelloff, G.J., Hylton, N.M., Berry, D.A., Esserman, L.J., 2009. I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clin. Pharmacol. Ther. 86 (1), 97–100.

Berry, D.A., 1987. Interim analysis in clinical trials: the role of likelihood principle. Am. Stat. 41 (2), 117–122.

Berry, D.A., 2004. Bayesian statistics and the efficiency and ethics of clinical trials. Statist. Sci. 19 (1), 175–187.

Berry, D.A., 2006. Bayesian clinical trials. Nat. Rev. Drug Discov. 5 (1), 27–36.

Berry, D.A., 2015a. Presentation at GBM AGILE workshop, August 11–12, Phoenix, AZ.

Berry, D.A., 2015b. The brave new world of clinical cancer research: adaptive biomarker-driven trials integrating clinical practice with clinical research. Mol. Oncol. 9 (5), 951–959.

Berry, D.A., Fristedt, B., 1985. Bandit Problems: Sequential Allocation of Experiments, first ed. In: Mongraphs on Statistics and Applied Probability, Springer Netherlands.

Center for Devices and Radiological Health of the U.S. Food and Drug Administration. Feb. 2010. Guidance for industry and FDA staff: Guidance for the use of Bayesian statistics in medical device clinical trials. http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf. (Accessed 14 March 2015).

Cheng, Y., Su, F., Berry, D.A., 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. J. Am. Stat. Assoc. 58 (302), 388–400.

DeGroot, M.H., 1970. Optimal Statistical Decisions. McGraw-Hill Book Company, New York.

Deley, M.C.L., Ballman, K.V., Marandet, J., Sargent, D., 2012. Taking the long view: how to design a series of Phase III trials to maximize cumulative therapeutic benefit. Clin. Trials 9 (3), 283–292.

Eichler, H.G., Abadie, E., Breckenridge, A., et al., 2011. Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response. Nat. Rev. Drug Discov. 10 (7), 495–506.

Eichler, H.G., Abadie, E., Raine, J.M., Salmonson, T., 2009. Safe drugs and the cost of good intentions. New Engl. J. Med. 360 (14), 1378–1380.

Eichler, H.G., Bloechl-Daum, B., Brasseur, D., et al., 2013. The risks of risk aversion in drug regulation. Nat. Rev. Drug Discov. 12 (12), 907–916.

Eichler, H.G., Oye, K., Baird, L.G., et al., 2012. Adaptive licensing: taking the next step in the evolution of drug approval. Clin. Pharmacol. Ther. 91 (3), 426–437.

Eichler, H.G., Pignatti, F., Flamion, B., Leufkens, H., Breckenridge, A., 2008. Balancing early market access to new drugs with the need for benefit/risk data: a mounting dilemma. Nat. Rev. Drug Discov. 7 (10), 818–826.

European Medicines Agency. Dec. 2010. Road map to 2015: The European medicines agencys contribution to science, medicines and health. http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/01/WC500101373.pdf. (Accessed 15 March 2015).

European Medicines Agency, 2014. European Medicines Agency. Adaptive pathways to patients: Report on the initial experience of the pilot project. Technical Report EMA/758619/2014.

Freedman, B., 1987. Equipoise and the ethics of clinical research. New Engl. J. Med. 317 (3), 141–145.

Friedman, L.M., Furberg, C.D., DeMets, D.L., 2010. Fundamentals of Clinical Trials: A Practical Approach, fourth ed. Springer, New York.

Greener, M., 2005. Drug safety on trial. EMBO Rep. 6 (3), 202–204.

Grieve, A.P., 2015. How to test hypotheses if you must. Pharm. Stat. 14 (2), 139–150.

Harrington, D., Parmagiani, G., 2016. I-SPY 2-a glimpse of the future of phase 2 drug development? New Engl. J. Med. 375, 7–9.

Jennison, C., Turnbull, B.W., 2010. Group Sequential Methods with Applications to Clinical Trials. CRC Press.

Lenert, L.A., Markowitz, D.R., Blaschke, T.F., 1993. Primum non nocere? Valuing of the risk of drug toxicity in therapeutic decision making. Clin. Pharmacol. Ther. 53 (3), 285–291.

Lo, A.W., 2017. Discussion: new directions for the FDA in the 21st century. Biostatistics 18 (3), 404–407.

McGlothlin, A.E., Lewis, R.J., 2014. Minimal clinically important difference: defining what really matters to patients. JAMA 312 (13), 1342–1343.

McNaughton, R., Huet, G., Shakir, S., 2014. An investigation into drug products withdrawn from the eu market between 2002 and 2011 for safety reasons and the evidence used to support the decision-making. BMJ Open 4 (1), e004221.

Montazerhodjat, V., Chaudhuri, S., Sargent, D., Lo, A.W., 2017. Use of bayesian decision analysis to minimize harm in patient-centered randomized clinical trials in oncology. JAMA Oncol. http://dx.doi.org/10.1001/jamaoncol.2017.0123.

Murray, C.J. L., Abraham, J., Ali, M.K., et al., 2013. The State of U.S. health, 1990–2010: burden of diseases, injuries, and risk factors. JAMA 310 (6), 591–608.

National Biomarker Development Alliance. 2017. GBM AGILE. http://nbdabiomarkers.org/gbm-agile. (Accessed 7 September 2017).

National Institute for Health and Care Excellence. Feb. 2008. Quality adjusted life years (QALYs) and severity of illness: Report 10. https://www.nice.org.uk/Media/Default/Get-involved/Citizens-Council/Reports/CCReport10QALYSeverity.pdf. (Accessed 9 July 2015).

Paavonen, J., Naud, P., Salmerón, J., et al., 2009. Efficacy of human papillomavirus (HPV)-16/18 AS04-adjuvanted vaccine against cervical infection and precancer caused by oncogenic HPV types (PATRICIA): final analysis of a double-blind, randomised study in young women. Lancet 374 (9686), 301–314.

Piantadosi, S., 2005. Clinical Trials: A Methodological Perspective, second ed. John Wiley & Sons, New York.

Pocock, S.J., 1983. Clinical Trials: A Practical Approach. Wiley, New York.

Polyak, K., 2011. Heterogeneity in breast cancer. J. Clin. Invest. 121 (10), 3786–3788.

ProCon.org, 2014. 35 FDA-Approved Prescription Drugs Later Pulled from the Market.

Salomon, J.A., Vos, T., Hogan, D.R., et al., 2012. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the global Burden of Disease Study 2010. Lancet 380 (9859), 2129–2143.

Selker, H.P., Oye, K.A., Eichler, H.G., et al., 2014. A proposal for integrated efficacy-to-effectiveness (E2E) clinical trials. Clin. Pharmacol. Ther. 59 (2), 147–153.

Spiegelhalter, D.J., Freedman, L.S., Parmar, M.K. B., 1994. Bayesian approaches to randomized trials. J. R. Stat. Soc. Ser. A Stat. Soc. 157 (3), 357–416.

Steuer, C., Papadimitrakopoulou, V., Herbst, R., Redman, M., F., Hirsch, Mack, P., Ramalingam, S., Gandara, D., 2015. Innovative clinical trials: The LUNG-MAP study. Clin. Pharmacol. Ther. 97, 488–491.

U.S. Congress. Apr. 1999. Title 21, Code of Federal Regulations, part 312, subpart E: Drugs intended to treat life-threatening and severely-debilitating illnesses. http://www.gpo.gov/fdsys/pkg/CFR-1999-title21-vol5/pdf/CFR-1999-title21-vol5-part312-subpartE.pdf. (Accessed 20 April 2014).

U.S. Congress. Jul. 2012. Food and Drug Administration Safety and Innovation Act, public law 112-144. http://www.gpo.gov/fdsys/pkg/PLAW-112publ144/pdf/PLAW-112publ144.pdf. (Accessed 20 April 2014).

U.S. Food and Drug Administration. Sep. 1998. Guidance for industry: E9 statistical principles for clinical trials. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf. (Accessed 30 December 2015).

U.S. Food and Drug Administration. Jan. 2006. Guidance for industry: Fast track drug development programs designation, development, and application review. http://www.fda.gov/downloads/Drugs/Guidances/ucm079736.pdf. (Accessed 20 April 2015).

U.S. Food and Drug Administration. Feb. 2013. Draft PDUFA V implementation plan: Structured approach to benefit-risk assessment in drug regulatory decision-making. http://www.fda.gov/downloads/ForIndustry/UserFees/PrescriptionDrugUserFee/UCM329758.pdf. (Accessed 20 April 2014). Fiscal Years 2013–2017.

U.S. Food and Drug Administration. Apr. 2013. Federal register notice. http://www.gpo.gov/fdsys/pkg/FR-2013-04-11/pdf/2013-08441.pdf. (Accessed 20 April 2014).

U.S. Food and Drug Administration. June 2013. Guidance for industry: Expedited programs for serious conditions— drugs and biologics. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM358301.pdf. (Accessed 20 April 2015).