NBER WORKING PAPER SERIES

IS THE FDA TOO CONSERVATIVE OR TOO AGGRESSIVE?: A BAYESIAN DECISION ANALYSIS OF CLINICAL TRIAL DESIGN

Vahid Montazerhodjat
Andrew W. Lo

Working Paper 21499
http://www.nber.org/papers/w21499

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2015

We thank Ernie Berndt, Don Berry, Bruce Chabner, Jayna Cummings, Mark Davis, Hans-Georg Eichler, Williams Ettouati, Gigi Hirsch, Tomas Philipson, and Nora Yang for helpful comments and discussion. 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, any of the individuals acknowledged above, or the National Bureau of Economic Research. Research support from the MIT Laboratory for Financial Engineering is gratefully acknowledged.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w21499.ack

Is the FDA Too Conservative or Too Aggressive?: A Bayesian Decision Analysis of Clinical Trial Design

Vahid Montazerhodjat and Andrew W. Lo

NBER Working Paper No. 21499
August 2015
JEL No. C11,C12,C44,I10,I12,I13,I18

ABSTRACT

Implicit in the drug-approval process is a 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 relative costs 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 27.9%. However, 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%. 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.

Vahid Montazerhodjat
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
montazer@mit.edu

Andrew W. Lo
MIT Sloan School of Management
100 Main Street, E62-618
Cambridge, MA 02142
and NBER
alo-admin@mit.edu

Contents

1 Introduction

Randomized clinical trials (RCTs) have been widely accepted as the most reliable approach for determining the safety and efficacy of drugs and medical devices [1, 2], and their outcomes largely determine whether new therapeutics are approved by regulatory agencies such as the U.S. Food and Drug Administration (FDA). Because RCTs often involve several thousand human subjects and require years to complete, the FDA is sometimes criticized for being too conservative, requiring trials that are "overly large" [3] and using too conservative a threshold of statistical significance.

In response to these concerns, the FDA has gone to great lengths to expedite the approval process for drugs intended to treat serious conditions and rare diseases [4, 5].[1] Four programs—fast-track, breakthrough-therapy, accelerated-approval, and priority-review designations—provide faster reviews and/or use surrogate endpoints to judge efficacy. However, the published descriptions [4, 5] do not indicate any difference 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. Hence, from the patient's perspective, the approval criteria in these programs may still seem too conservative, especially for terminal illnesses with no existing treatment options. Moreover, a large number of compounds are not eligible for these special designations, and some physicians have argued that the regulatory safety requirements for drugs targeting non-cancer life-threatening diseases, e.g., cirrhosis of the liver and hypertensive heart disease, should be relaxed.

At the heart of this debate is the unavoidable regulatory trade-off between maximizing the benefits of effective therapies to patients and minimizing the risk to those who do not respond to such therapies. Even under the current thresholds of statistical significance, both the U.S. and Europe have seen harmful drugs with severe side effects make their way into the market [6–9]. Therefore, the FDA and the European Medicine Agency (EMA)—government agencies mandated to protect the public—are understandably reluctant to employ more risk-tolerant or aggressive statistical criteria to judge the efficacy of a drug. However, we show in this article that when the risk of adverse side effects is explicitly weighed against the severity of the disease, the standard thresholds of statistical significance are sometimes more

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

conservative than the derived optimal significance levels, and substantially so for the most serious afflictions such as pancreatic cancer. On the other hand, the same conventional statistical thresholds are more aggressive than the optimal thresholds for milder illnesses such as prostate cancer. Therefore, criticizing drug regulatory agencies for being overly conservative or aggressive without explicitly specifying the burden of disease, i.e., the therapeutic costs and benefits for current and future patients, is uninformed and vacuous.

In statistical terms, regulators must weigh the cost of a Type I error—approving an ineffective therapy—against the cost of a Type II error—rejecting an effective therapy. However, the term "cost" in this context refers not just to direct financial costs, but also includes the consequences of incorrect decisions for all current and future patients. Complicating this process is the fact that these trade-offs 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 (risks) of the 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 a publicly traded pharmaceutical company whose shareholders will bear the enormous cost of wrongful death litigation.

The need to balance these competing considerations in decision-making for drug-approval has long been recognized by clinicians, drug-regulatory experts and other stakeholders [10–12]. It has also been recognized that these competing factors should be taken into account when designing clinical trials [13–15] and one approach to quantify this need is to assign different costs to the different outcomes [15].

In this paper, we explore these trade-offs explicitly by applying a Bayesian decision analysis (BDA) framework to the design of RCTs as advocated by [15, 16]. In this framework, Type I and II errors are assigned different costs, as first suggested by [13–15], 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 concept of assigning costs to outcomes and employing cost-minimization techniques

to determine optimal decisions is well known [17]. 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 burden-of-disease data. This approach yields a systematic, objective, transparent, and repeatable process for making regulatory decisions that reflects differences in disease-specific parameters. Moreover, given a specific statistical threshold, and assuming that this threshold is optimal from a BDA perspective, we can invert the relationship between cost parameters and their corresponding BDA-optimal tests to impute the costs implicit in a given clinical trial design. This allows us to infer the FDA's implicit weighting of Type I and II errors, which yields an objective measure of whether its approval thresholds are too conservative or aggressive.

Using U.S. Burden of Disease Study 2010 data [18], we show that the current standards of drug-approval are weighted more on avoiding a Type I error (approving ineffective therapies) rather than a Type II error (rejecting effective therapies). For example, the standard Type I error of 2.5% is considerably more conservative than the BDA-optimal Type I error of 27.9% 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 3 February 2013). The BDA-optimal size for these clinical trials is 27.9%, reflecting the fact that, for these desperate patients, the cost of trying an ineffective drug is considerably 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% 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 = 16.6%) and hypertensive heart disease (optimal size = 8.1%).

Although there are obvious utilitarian reasons for weighting Type I errors more heavily, they do not necessarily apply to all diseases or stakeholders. For terminal illnesses where patients have no choice but death, the relative costs of Type I and II errors are very different than for non-life-threatening conditions. This difference is clearly echoed in the Citizens Council report published by the U.K.'s National Institute for Health and Care Excellence (NICE) [19], and has also been documented in a series of public meetings held by the FDA as part of its five-year Patient-Focused Drug Development Program, in which the gap between patients' risk/benefit perception and the FDA's was apparent [20, 21]. Our BDA framework

incorporates the severity of the disease into its design—as advocated in part 312, subpart E of title 21 Code of Federal Regulation (CFR) [22]—and the FDA reports [20, 21], among many other sources, can be used to determine the relative cost parameters from the patients' and even the general public's perspective in an objective and transparent manner. As suggested in [23], using hard evidence, i.e., available data, for assigning costs to different events is a feasible remedy to the controversy often surrounding Bayesian techniques due to their subjective judgment factor in the cost-assignment process. In fact, Bayesian techniques have survived controversy and are currently used extensively in clinical trials for medical devices, mainly 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 [23].

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. The BDA framework is introduced in Section 4, which can be shown to mitigate the shortcomings of the classical approach, and the BDA-optimal fixed-sample test is then derived. We apply this framework in Section 5 by first estimating the parameters of the Bayesian model using the U.S. Burden of Disease Study 2010 [18]. Using these estimates, we 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 conclude in Section 7.

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 for maximizing the approval probability for an effective drug is obvious. In the classical (frequentist) approach to hypothesis testing—currently the standard 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 of the test. Type II error occurs when an effective drug is rejected, and the complement of the probability of this error 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 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, and, by choosing a large enough sample size, it maintains a power for the alternative hypothesis, right around another somewhat arbitrary level, usually $1 - \beta = 80\%$.

The arbitrary nature of these values for the size and power of the test raises legitimate questions about their justification. As will be seen later, these particular values correspond to a specific situation, which need not (and most likely does not) apply to clinical trials employed to test new drugs for different diseases. It is also worth noting that these numbers were brought to the design paradigm of clinical trials from other industries, in particular, the manufacturing industry. Therefore, it is reasonable to ask if these totally 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 must be much different from the results of mistakenly rejecting an effective drug for many patients with a life-threatening disease, who may desperately be looking for effective therapeutics. In other words, there must be different *costs* associated with each of these wrong rejections.

In addition to the arbitrary nature of the commonly used values for the size and power of tests, there is an important ethical issue with regard to 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 dismisses that *at least* half of the recruited subjects are exposed to ineffective treatment during the trial, assuming a balanced two-arm RCT [15, 24]. This ethical issue, along with financial considerations, is the principal reason that the sample size in classical trial design is not further increased to get more 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 classical approach to the design of clinical trials does not take into account the possible 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 this and similar arguments, 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 in [14, 24, 25]. We refer to the population to be affected by the outcome of the trial as *the target population* in the rest of this paper, and note that it is the same as the patient horizon originally proposed in [13, 14] and later used in [24, 25]. 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.

3 A Review of RCT Statistics

In this section, we explain the basic statistics of RCTs and define the notation employed in this paper. We begin with the design of the balanced two-arm RCT where the subjects are randomly assigned to either the treatment or control arm, and there is an equal number of subjects in each arm. 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 making any observations. Furthermore, only after collecting *all* the observations, shall a decision be made on whether or not the drug is effective. However, our approach is equally applicable to more sophisticated designs since the more novel designs usually try to mimic the statistical performance of a fixed-sample test, e.g., frequentist power and size, while minimizing sample size.

A quantitative primary endpoint is assumed for the trial. 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 [2, 26]. The subjects in the treatment and control arms receive the drug and placebo, respectively, and each subject's response is 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,

instead of the placebo, is assumed to be administered to the patients in the control arm. In either situation, it is natural to assume that the administered drug to the control arm patients is not toxic. The response variables in the treatment arm, denoted by $\{T_1, \ldots, T_n\}$, are independent and identically distributed (iid), where $T_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_t, \sigma^2)$. Similarly, for the control (placebo) arm responses, represented by $\{P_1, \ldots, P_n\}$, we assume $P_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_p, \sigma^2)$, 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.

Furthermore, we focus only on superiority trials, in which the drug candidate is likely to have either a positive effect or no effect (possibly with adverse side effects).[2] Let us define the treatment effect of the drug, δ, as the difference of the response means in the two arms, i.e., $\delta \triangleq \mu_t - \mu_p$. The event in which the drug is ineffective and has adverse side effects defines our null hypothesis, H_0, corresponding to $\delta = 0$ (and 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, 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 [26]. The Z-statistic, Z_n, is then compared to a critical value, λ_n, and the null hypothesis is not rejected, denoted by $\widehat{H} = H_0$, if the Z-statistic is smaller than the critical value. Otherwise, the null hypothesis is rejected, represented by $\widehat{H} = H_1$:

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

[2]Non-inferiority trials—where 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.

As is 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, where the significance level is exogenous and independent of the sample size. Since a fixed-sample test is completely characterized 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 the sake of 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

In the following, we propose a quantitative framework to explicitly take into account the severity of the disease when determining the sample size and critical value of a fixed-sample test. We first define 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 critical value for the test are then jointly determined to minimize the expected cost of the trial. As stated in Section 1, the term "cost" in this paper refers to the health consequences of incorrect decisions for all current and future patients, and not necessarily the financial cost.

Our methods are similar to [14], although the cost model used here is different from his. Furthermore, the authors of [25] 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.

Cost Model

The costs associated with a clinical trial can be categorized into two groups: in-trial costs and post-trial costs, where 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, 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 the *Type II cost*.

Specifying asymmetric costs for Type I and Type II errors allows 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, where 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, e.g., mild anemia or secondary infertility, the cost of adverse side effects can be much larger than the cost of not approving an effective drug for the disease, hence, the Type I cost can be much larger than the Type II cost, i.e., $C_1 \gg C_2$. If the severity of the disease is intermediate, e.g., moderate anemia or mild dementia, then these two post-trial costs may be more or less the same, i.e., $C_1 \approx C_2$.

Furthermore, 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 in favor of/against the null hypothesis; therefore, the larger the values of C_1 and C_2. Let us assume this relation is linear in the target population size. More precisely, if the size of the target population is N, assume there exist two constants, c_1 and c_2, which are independent of the disease prevalence and depend only on the adverse side effects 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.

	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	$n\gamma C_2$

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

In-trial costs are mainly related to patients' exposure to inferior treatment, e.g., the exposure of enrolled patients to an ineffective but toxic 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, since there are n subjects in the treatment arm taking this drug, they 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 not to be toxic. However, if the drug is effective, the situation is quite different. In this case, for every additional patient in the trial, there will be an incremental delay in the emergence of the drug in the market. This delay affects all patients, both inside or 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 $\gamma = 4 \times 10^{-5}$). All the cost categories associated with a fixed-sample test are tabulated in Table 1.

Now, 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 define the incurred cost, denoted by $C(H, Z_n, \texttt{fxd}(n, \lambda_n))$, as the following:

$$C(H, Z_n, \texttt{fxd}(n, \lambda_n)) = \begin{cases} Nc_1 \mathbb{1}_{\{Z_n \geq \lambda_n\}} + nc_1, & H = H_0 \\ \\ Nc_2 \mathbb{1}_{\{Z_n < \lambda_n\}} + n\gamma N c_2, & H = H_1 \end{cases}, \tag{4}$$

where $\mathbb{1}$ is the indicator function and 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, caused by Type I error, i.e., approving the ineffective drug, which yields 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 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:

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

where Φ is the cumulative distribution function of a standard normal random variable, $Z \sim \mathcal{N}(0,1)$, and E is the expectation operator. It is worth noting that if $p_0 = p_1 = 0.5$, then $\overline{c}_2 = \frac{p_1 c_2}{p_0 c_1}$ reduces to $\overline{c}_2 = \frac{c_2}{c_1}$, i.e., the normalized Type II cost. For the remainder of this paper, we assume a non-informative prior, i.e., $p_0 = p_1 = 0.5$, and hence regard $\overline{c}_2$ as the normalized Type II cost that is the ratio of Type II cost to Type I cost.

A non-informative prior is consistent with the "equipoise" principle of two-arm clinical trials [27]. However, in some cases we can formulate more informed priors based on information accumulated through earlier-phase trials and other sources. In such cases, the randomization of patients should reflect this information—especially, for life-threatening conditions—for ethical reasons, and the natural framework for doing so is a Bayesian adaptive design [3, 28]. Although this framework is beyond the scope of our current analysis, BDA can easily be applied to adaptive designs and we will consider this case in future research.

The optimal sample size n^* and critical value λ_n^* are determined such that the expected cost of the trial, given by (5), is minimized (see Appendix A.1 for a detailed description). The fixed-sample test with these two parameters, i.e., $\texttt{fxd}(n^*, \lambda_n^*)$, will be referred to as the BDA-optimal fixed-sample test. Furthermore, given any fixed-sample test, $\texttt{fxd}(n, \lambda)$—and assuming the test is a BDA-optimal test for a disease with unknown severity (Type II cost) and prevalence—we can impute the severity of disease and its prevalence (see Appendix A.2) implied by the threshold λ.

5 Estimating the Cost of Disease

In this section, the two cost parameters, c_1 and c_2, associated with adverse effects of medical treatment and severity of the disease to be treated, respectively, are estimated. To estimate these two parameters, we use the U.S. Burden of Disease Study 2010 [18], which follows the same methodology as of the comprehensive Global Burden of Disease Study 2010 (GBD 2010), however, with only U.S.-level data. Since only the ratio of c_2 over c_1, i.e., $\overline{c}_2$, appears in the expected cost of the trial in (5), we use the severity estimates of adverse effects of medical treatment and of disease in the U.S. for c_1 and c_2, respectively.

One of the key factors in quantifying the burden of disease and loss of health due to different diseases and injuries 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 heath, 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 [29], 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, because c_2 is the severity of the disease to be treated, taking into account the current state of available therapies for the disease. We estimate the overall

severity of disease using the following equation:

$$c_2 = \frac{D + \text{YLD}}{D + N}, \tag{6}$$

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; hence, 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 people afflicted with, or who died from, 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 get 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) [30]. This technique facilitates meaningful comparison of rates for different populations and diseases.

To estimate c_1, which is the current cost of adverse effects of medical treatment per patient, we insert the corresponding numbers for the adverse effect of medical treatment in the U.S. from the U.S. Burden of Disease Study 2010 [18] into (6), and the result is $c_1 = 0.07$. The value of c_1 can be made more precise and tailored to the drug candidate being tested if the information from earlier clinical phases, e.g., Phase I and Phase II, is used. However, for simplicity, we only consider a common value for c_1 for all diseases.

6 BDA-Optimal Tests for the Most Deadly Diseases

Using (6) and the YLD, death and prevalence rates reported in the U.S. Burden of Disease Study 2010 [18], we can now estimate the severity of some of the leading causes of death in the U.S. in 2010. Using the estimated severity of each disease, we can then determine the BDA-optimal fixed-sample test for a drug intended to treat that disease. The drug is assumed to have either a positive effect on the disease (corresponding to $\delta_0 = \frac{\sigma}{8}$) or no effect with adverse side effects (corresponding to $\delta = 0$).

The leading causes of death, listed in Table 2, are determined in [18] by ranking diseases

and injuries based on their associated YLLs (Years of Life Lost due to premature death) in the U.S. in 2010. The following categories, while among the leading causes of premature mortality in the U.S., are omitted from Table 2 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 violence (12), preterm birth complications (14), drug-use disorders (15), other cardiovascular/circulatory diseases (17), congenital anomalies (19), poisonings (26), and falls (29). We have also subdivided two categories into subcategories in Table 2: stroke is listed as ischemic stroke (3a) and non-ischemic stroke (3b), and lower respiratory tract infections is divided into four diseases (11a)–(11d). These choices yield 25 leading causes of death for which we compute BDA-optimal thresholds and compare them to more traditional values.

The estimated severity for each disease, c_2, is reported in the fourth column of Table 2. As can be seen, some cancers are not quite as severe as other non-cancerous diseases. For instance, prostate cancer ($c_2 = 0.05$), is much less harmful than cirrhosis ($c_2 = 0.49$), 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 $c_2 = 0.71$. Using this measure of severity, we have an objective data-driven framework where different diseases with different afflicted populations can be compared with one another.

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 listed in Table 2. The sample size, critical value, size, and statistical power of these BDA-optimal tests are reported in Table 2. For comparison, we have also listed the imputed prevalence and severity for three conventional 2.5%-level fixed-sample tests in the last three rows of Table 2 under the assumption that these conventional thresholds are BDA-optimal (see Appendix A.2).

Some of the diseases listed in Table 2 are no longer a single disease but rather a collection of diseases with heterogeneous biological and genetic profiles, and with distinct patient populations [31, 32], 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 [32]. However, because burden-of-disease data

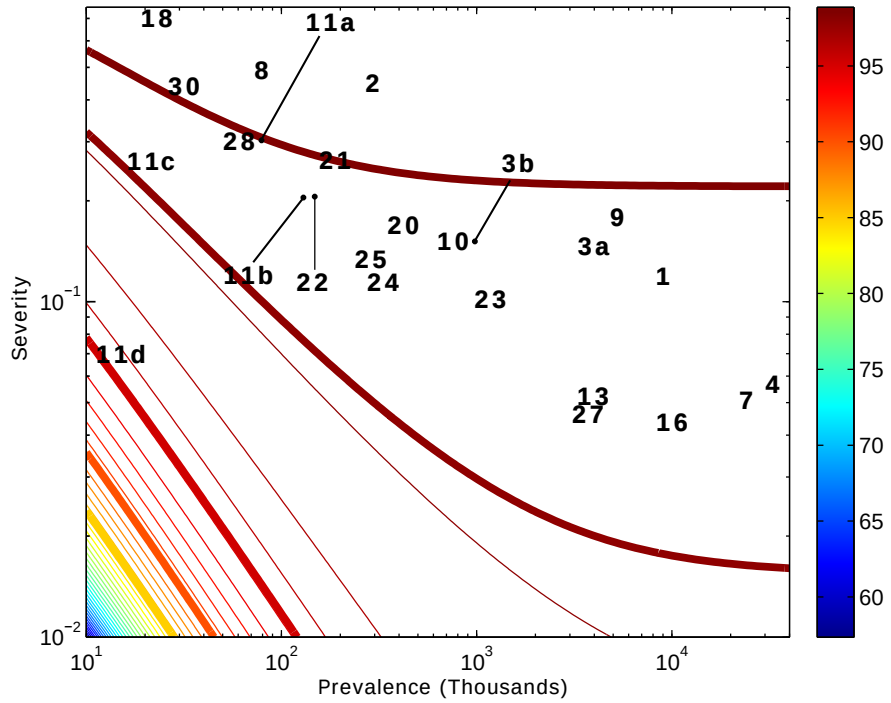
YLL Rank	Disease Name	Prevalence (Thousands)	Severity	Optimal Sample Size	Optimal Critical Value	Size (%)	Power (%)
1	Ischemic heart disease	8,895.61	0.12	2,028	1.845	3.25	98.36
2	Lung cancer	289.87	0.45	1,373	1.055	14.56	98.68
3a	Ischemic stroke	3,932.33	0.15	1,936	1.744	4.06	98.40
3b	Hemorrhagic/other non-ischemic stroke	949.33	0.16	1,902	1.709	4.37	98.40
4	Chronic obstructive pulmonary disease	32,372.11	0.06	2,343	2.177	1.47	98.22
7	Diabetes	23,694.90	0.05	2,387	2.221	1.32	98.20
8	Cirrhosis of the liver	78.37	0.49	1,300	0.969	16.64	98.67
9	Alzheimer's disease	5,145.03	0.18	1,845	1.640	5.05	98.45
10	Colorectal cancer	798.90	0.15	1,905	1.714	4.33	98.40
11a	Pneumococcal pneumonia	84.14	0.30	1,550	1.311	9.49	98.49
11b	Influenza	119.03	0.20	1,744	1.552	6.03	98.38
11c	H influenzae type B pneumonia	21.15	0.26	1,453	1.279	10.04	98.17
11d	Respiratory syncytial virus pneumonia	14.90	0.07	1,491	1.692	4.53	95.73
13	Breast cancer	3,885.25	0.05	2,374	2.212	1.35	98.19
16	Chronic kidney disease	9,919.02	0.04	2,447	2.283	1.12	98.17
18	Pancreatic cancer	22.67	0.71	1,027	0.587	27.86	98.76
20	Cardiomyopathy	416.31	0.17	1,853	1.659	4.86	98.41
21	Hypertensive heart disease	185.26	0.27	1,633	1.401	8.06	98.50
22	Leukemia	139.75	0.21	1,724	1.522	6.40	98.41
23	HIV/AIDS	1,159.58	0.10	2,087	1.915	2.77	98.31
24	Kidney cancers	328.94	0.12	2,011	1.846	3.24	98.29
25	Non-Hodgkin lymphoma	282.94	0.13	1,944	1.772	3.82	98.32
27	Prostate cancer	3,709.70	0.05	2,414	2.252	1.22	98.17
28	Brain and nervous system cancers	59.76	0.30	1,524	1.290	9.86	98.46
30	Liver cancer	31.27	0.44	1,302	1.004	15.77	98.56
—	2.5%-level Fixed-Sample (85% power)	15.12	0.02	1,150	1.960	2.50	85.02
—	2.5%-level Fixed-Sample (90% power)	17.51	0.02	1,345	1.960	2.50	90.00
—	2.5%-level Fixed-Sample (95% power)	24.60	0.04	1,664	1.960	2.50	95.01

Table 2: 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 as well as 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.

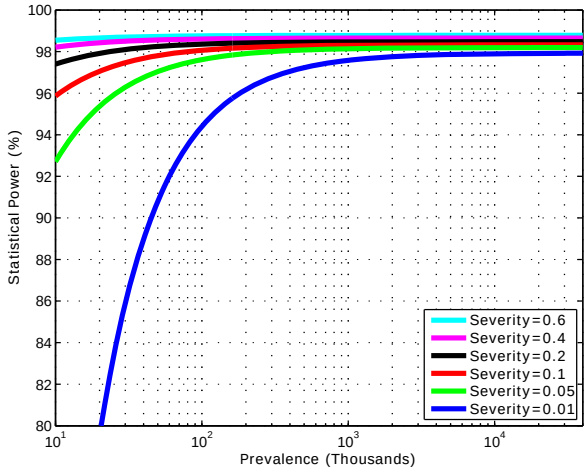
are not yet available for these subdivisions, we use the conventional categories in Table 2, i.e., where each cancer type is decided based on the organ host of the tumor.

The reported values for the power of BDA-optimal tests are quite high (all but one have power larger than 98%). This is because the overall burden of disease ($C_2 = Nc_2$) associated with each of these diseases is quite high, due to either severity (large c_2), e.g., pancreatic cancer, or high prevalence (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 (c_2), and many cancers are being reclassified as orphan diseases through the use of biomarkers and personalized medicine [32]. Therefore, not approving an effective drug is a costly option by this measurement, hence these BDA-optimal tests exhibit high power to detect positive treatment effects. This general dependence of the statistical power on the overall burden of disease, i.e., its prevalence multiplied by its severity, can be observed in Figure 1. In Figure 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., $Nc_2 = cte$, which are straight lines with negative slope on a log-log graph. Also, to facilitate visualizing where each disease in Table 2 lies in the prevalence-severity plane, we have superimposed the YLL rank of each disease in Figure 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 Figures 1(b) and (c).

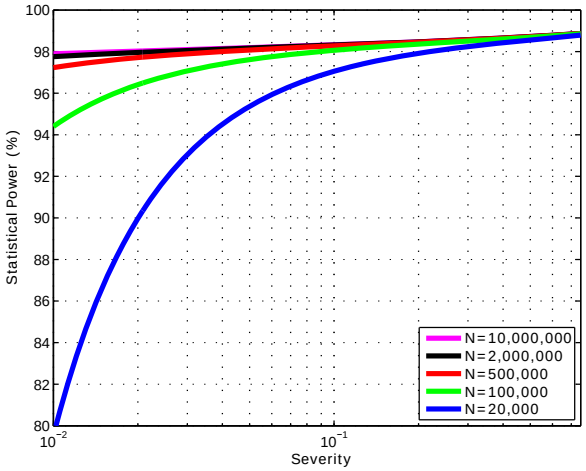
In sharp contrast to the consistently high power for the BDA-optimal tests in Table 2, the size of these tests varies dramatically across different diseases. As is seen in Table 2, with few exceptions, the size of the test mainly depends on the severity of the disease. In general, as the severity of 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 0.587, while for the conventional 2.5%-level fixed-sample test it is 1.960. This results in a relatively high size (27.86%) for the BDA-optimal test for a drug intended to treat pancreatic cancer, consistent with the necessity for greater willingness to approve drugs intended to treat life-threatening diseases that have no existing effective treatment.



(a)



(b)



(c)

Figure 1: The statistical power of the BDA-optimal fixed-sample test at the alternative hypothesis. 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 $1-\beta = 85\%, 90\%, 95\%, 98\%$, and 98.5% are highlighted in panel (a). The superimposed numbers in panel (a) denote the YLL rank of each disease in Table 2. 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.

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 1.32%. The size of BDA-optimal tests for a large range of severity and prevalence values is presented in Figure 2. The size monotonically increases with the severity of disease for any given prevalence, and as seen in Figures 2(a) and (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 Figure 2(a). This insensitivity of the size to the prevalence of disease makes our model quite robust against estimation noise in the disease prevalence.

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 Figure 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 Figure 3(c): as the severity of the disease increases, for a large enough target population ($N > 500{,}000$), the optimal sample size continuously shrinks to avoid any delay in getting 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, hence being exposed to 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 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.

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-

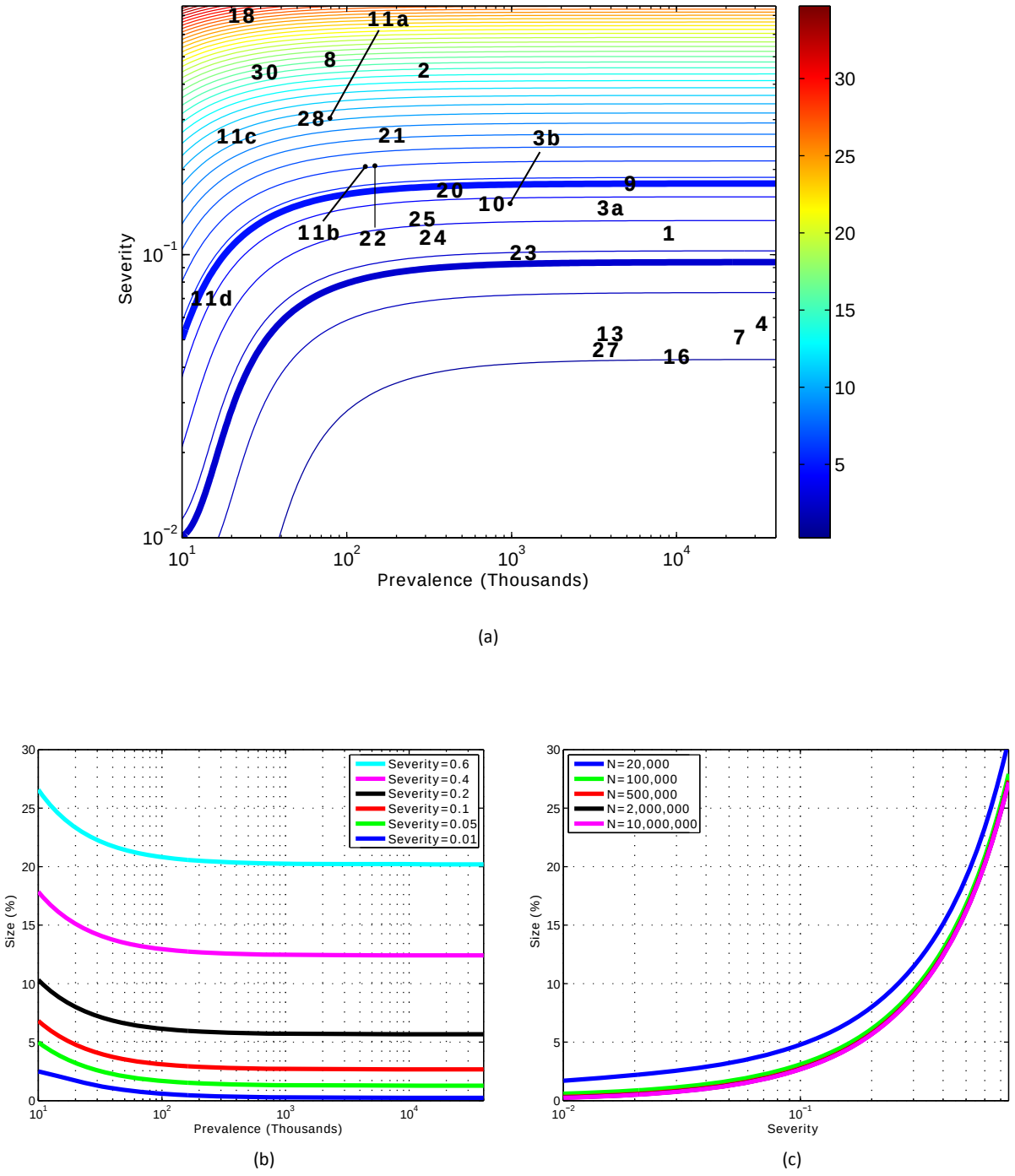


Figure 2: The size of the BDA-optimal fixed-sample test as a function of disease severity and prevalence. Panel (a) shows the contour levels for the size, while panels (b) and (c) demonstrate its cross-sections along the two axes. The contour lines corresponding to $\alpha = 2.5\%$ and $\alpha = 5.0\%$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

optimal tests is quite insensitive to the disease prevalence for large target populations (hence, horizontal contour lines in Figure 3(a) over large values of prevalence), which suggests that these results are robust.

Finally, inspecting the conventional fixed-sample tests and the disease prevalence and severity implied by them in Table 2 highlights the conservatism of current regulatory requirements imposed on clinical trials and their conduct if we assume that these values are BDA-optimal (see Appendix A.2).

7 Conclusion

To address the inflexibility of traditional frequentist designs for clinical trials, we explore the design of 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. Assuming that the current FDA standards represent BDA-optimal tests, the imputed costs implicit in these standards are overly conservative for the most deadly diseases and overly aggressive for the mildest ones. Therefore, changing the one-size-fits-all statistical criteria for FDA drug approval is likely to yield greater benefits to a greater portion of the population.

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 [20]. 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 easily 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 preserves important flexibility for regulatory decision-makers.

In fact, a Bayesian approach is ideally suited for weighing and incorporating patient perspectives into the drug-approval process. The 2012 Food and Drug Administration Safety and Innovation Act (FDASIA) [33] has "recognized the value of patient input to the entire drug development enterprise, including FDA review and decision-making." One proposal

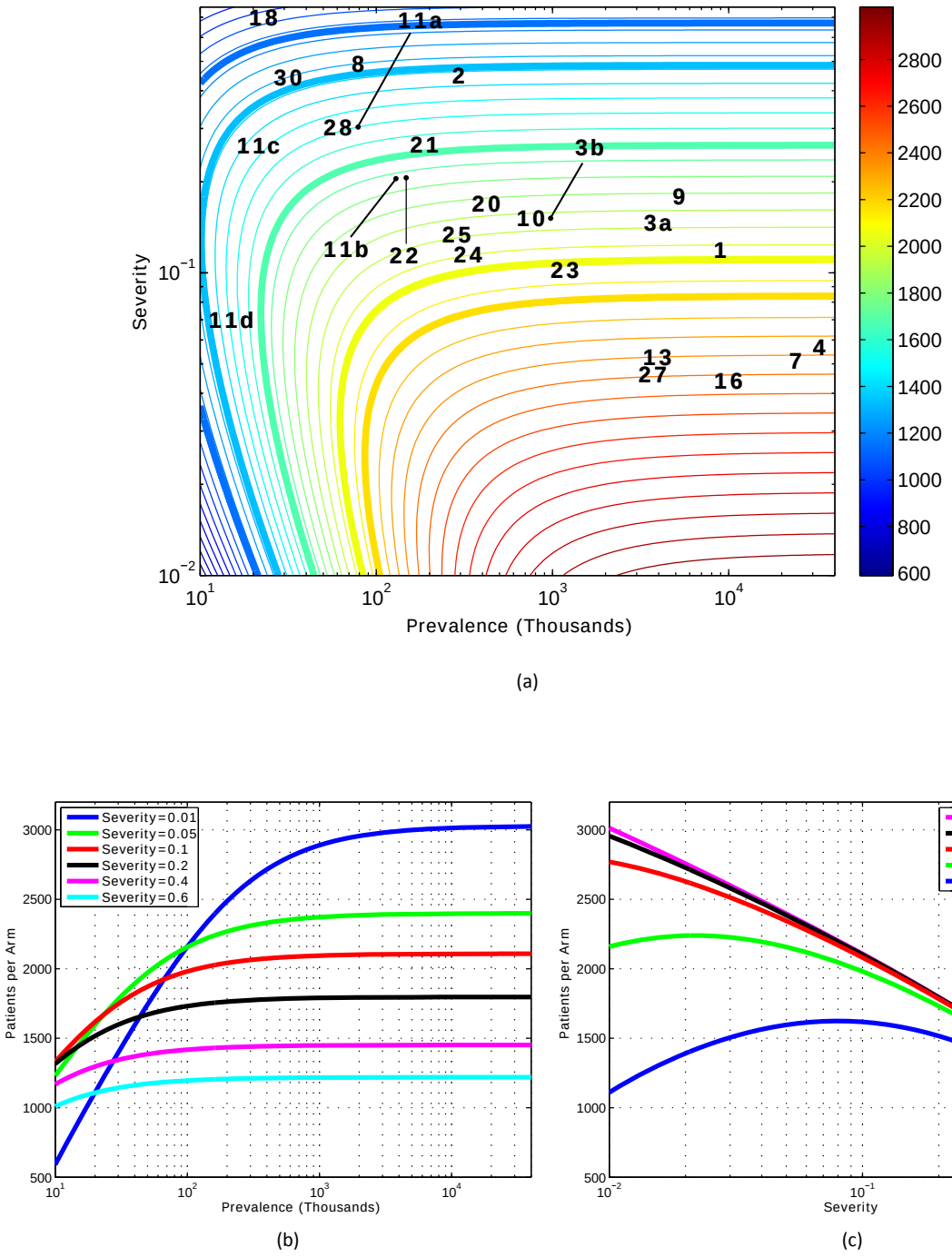


Figure 3: The sample size of the BDA-optimal fixed-sample test for different severity and prevalence values. Panel (a) shows the contour levels for the size, while panels (b) and (c) demonstrate its cross-sections along the two axes. The contour lines associated with the sample size of conventional fixed-sample tests with $\alpha = 2.5\%$ and $1 - \beta = 85\%, 90\%, 95\%, 98\%,$ and 98.5% are highlighted in panel (a). The superimposed numbers in panel (a) denote the YLL rank of each disease in Table 2. YLL: Number of years of life lost due to premature mortality, BDA: Bayesian decision analysis.

for implementing this aspect of FDASIA 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 an 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 such expanded composition, this advisory board could play an even broader role than the concept of a Citizens Council adopted by NICE.[3] The diverse set of stakeholders can provide crucial input to the FDA/EMA, reflecting the overall view of society on critical cost parameters. However, the role of such a committee should 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.

In fact, 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. Therefore, the FDA is in the unique role of formulating 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 should be industry neutral in its review process, nevertheless, ignoring scientific information in favor of 50/50 does not necessarily serve any stakeholder's interest. Moreover, using 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%, the size of the BDA-optimal test would be 51.2% rather than 27.9%, leading to many more approvals of such therapies. The BDA framework can yield decisions

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

that are both more economically efficient and more humane.

Finally,the drug-approval process is not always a binary choice, and in such cases, the BDA framework can be extended by defining costs for a finer set of events. In fact, 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 [34] (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 [35].

Moreover, new paradigms have also been proposed to address the risk associated with the binary nature of the current approval process, e.g., staggered approval [36, 37] and adaptive licensing [38], which the EMA is actively pursuing [39]. In fact, one of the design principles called for by [38] is less stringent statistical significance levels to be employed in efficacy trials for drugs targeting life-threatening diseases and/or rare conditions. Our BDA framework provides an explicit quantitative method for implementing this principle. The fact that the adaptive pathway has great potential to benefit all key stakeholders [40] provides more motivation for employing BDA in the drug-approval process.

References

[1] S. J. Pocock. *Clinical Trials: A Practical Approach*. Wiley, New York, 1983.

[2] L. M. Friedman, C. D. Furberg, and D. L. DeMets. *Fundamentals of Clinical Trials: A Practical Approach*. Springer, New York, 4th edition, 2010.

[3] D. A. Berry. Bayesian clinical trials. *Nat Rev Drug Discov*, 5(1):27–36, Jan. 2006.

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

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

[6] M. Greener. Drug safety on trial. *EMBO Rep*, 6(3):202–204, Mar. 2005.

[7] J. K. Aronson. Drug withdrawals because of adverse effects. In J.K. Aronson, editor, *A worldwide yearly survey of new data and trends in adverse drug reactions and interactions*, volume 30 of *Side Effects of Drugs Annual*, pages xxxi–xxxv. Elsevier, 2008. doi: http://dx.doi.org/10.1016/S0378-6080(08)00064-0.

[8] R. McNaughton, G. Huet, and S. Shakir. 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, Jan. 2014.

[9] ProCon.org. 35 FDA-approved prescription drugs later pulled from the market. Jan. 2014.

[10] L. A. Lenert, D. R. Markowitz, and T. F. Blaschke. Primum non nocere? valuing of the risk of drug toxicity in therapeutic decision making. *Clin Pharmacol Ther*, 53(3): 285–291, Mar. 1993.

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

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

[13] F. J. Anscombe. Sequential medical trials. *J Am Stat Assoc*, 58(302):365–383, 1963.

[14] T. Colton. A model for selecting one of two medical treatments. *J Am Stat Assoc*, 58 (302):388–400, 1963.

[15] D. A. Berry. Interim analysis in clinical trials: The role of likelihood principle. *Am Stat*, 41(2):117–122, May 1987.

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

[17] Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill Book Company, New York, 1970.

[18] C. J. L. Murray, J. Abraham, M. K. Ali, and et al. The state of U.S. health, 1990–2010: Burden of diseases, injuries, and risk factors. *JAMA*, 310(6):591–608, Jul. 2013.

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

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

[21] U.S. Food and Drug Administration. Federal register notice. Accessed April 20, 2014 at `http://www.gpo.gov/fdsys/pkg/FR-2013-04-11/pdf/2013-08441.pdf`, Apr. 2013.

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

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

[24] D. A. Berry. Bayesian statistics and the efficiency and ethics of clinical trials. *Stat Sci*, 19(1):175–187, 2004.

[25] Y. Cheng, F. Su, and D. A. Berry. Choosing sample size for a clinical trial using decision analysis. *Biometrika*, 90(4):923–936, Dec. 2003.

[26] C. Jennison and B. W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. CRC Press, 2010.

[27] B. Freedman. Equipoise and the ethics of clinical research. *New Engl J Med*, 317(3): 141–145, Jul. 1987.

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

[29] J. A. Salomon, T. Vos, D. R. Hogan, and et al. 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, Dec. 2012.

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

[31] K. Polyak. Heterogeneity in breast cancer. 121(10):3786–3788, Oct. 2011.

[32] D. A. Berry. The Brave New World of clinical cancer research: Adaptive biomarker-driven trials integrating clinical practice with clinical research. *Mol Oncol*, 9(5):951–959, Mar. 2015.

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

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

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

[36] H. G. Eichler, F. Pignatti, B. Flamion, H. Leufkens, and A. Breckenridge. 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, Oct. 2008.

[37] European Medicines Agency. Road map to 2015: The European medicines agency's contribution to science, medicines and health. Accessed March 15, 2015 at `http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/01/WC500101373.pdf`, Dec. 2010.

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

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

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

A Appendix

In this Appendix, we derive the expected-cost-minimizing critical value and sample size in Section A.1, and in Section A.2 we show how to impute the costs of Type I and Type II errors implicit in any one-sided fixed-sample test of a given size and power under the assumption that it is BDA-optimal.

A.1 Expected Cost Optimization

We determine the optimal sample size and the critical value for the fixed-sample test, $\texttt{fxd}(n, \lambda_n)$, by minimizing its expected cost in (5) over all possible values for n and λ_n. Keeping the sample size n fixed, the critical value λ_n that minimizes the expected cost, $C(\texttt{fxd}(n, \lambda_n))$ in (5), can be determined by setting the partial derivative of the expected cost, with respect to λ_n, to zero:

$$\frac{\partial}{\partial \lambda_n} C(\texttt{fxd}(n, \lambda_n))\bigg|_{\lambda_n = \lambda_n^*} = N p_0 c_1 \left[-\phi\left(-\lambda_n^*\right) + \overline{c}_2 \phi\left(\lambda_n^* - \delta_0 \sqrt{\mathcal{I}_n}\right)\right] = 0, \tag{7}$$

where ϕ is the probability density function of a standard normal random variable, i.e., $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$. Now, solving (7) for λ_n^* yields:

$$\lambda_n^* = -\frac{1}{\delta_0 \sqrt{\mathcal{I}_n}} \log\left(\overline{c}_2\right) + \frac{\delta_0 \sqrt{\mathcal{I}_n}}{2}, \tag{8}$$

where log is the natural logarithm and $\mathcal{I}_n = \frac{n}{2\sigma^2}$. By calculating the second derivative of the expected cost in (5), it is straightforward to prove that λ_n^*, given by (8), indeed minimizes the expected cost, $C(\texttt{fxd}(n, \lambda_n))$. Assuming $p_0 = p_1 = 0.5$, if Type I and Type II costs were equal, it is clear that the optimal critical value should be the midpoint of the means of the Z-statistic under the two hypotheses, hence the existence of the term $\frac{\delta_0 \sqrt{\mathcal{I}_n}}{2}$ in (8). However, in a general case, where the two costs are distinct, the first term in (8) plays the role of a correction term, and adjusts the optimal critical value to incorporate the difference between Type I and Type II costs.

Given a specific value of $\overline{c}_2$, the optimal critical value, i.e., λ_n^* in (8), can be considered a function of the sample size. The behavior of this function over different sample sizes for three values $c_2 = 0.01, 0.07, 0.34$, corresponding to $\overline{c}_2 = 0.2, 1, 5$, respectively, is depicted in

Figure 4, where the alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{8}$. The conventional critical value, regularly used for one-sided tests, i.e., $z_\alpha = \Phi^{-1}(1-\alpha) = 1.96$ for $\alpha = 2.5\%$, is also drawn in Figure 4 for comparison. It is observed that, in all of these cases, the optimal critical value changes with the sample size contrary to the classical critical value, which is independent of the sample size.

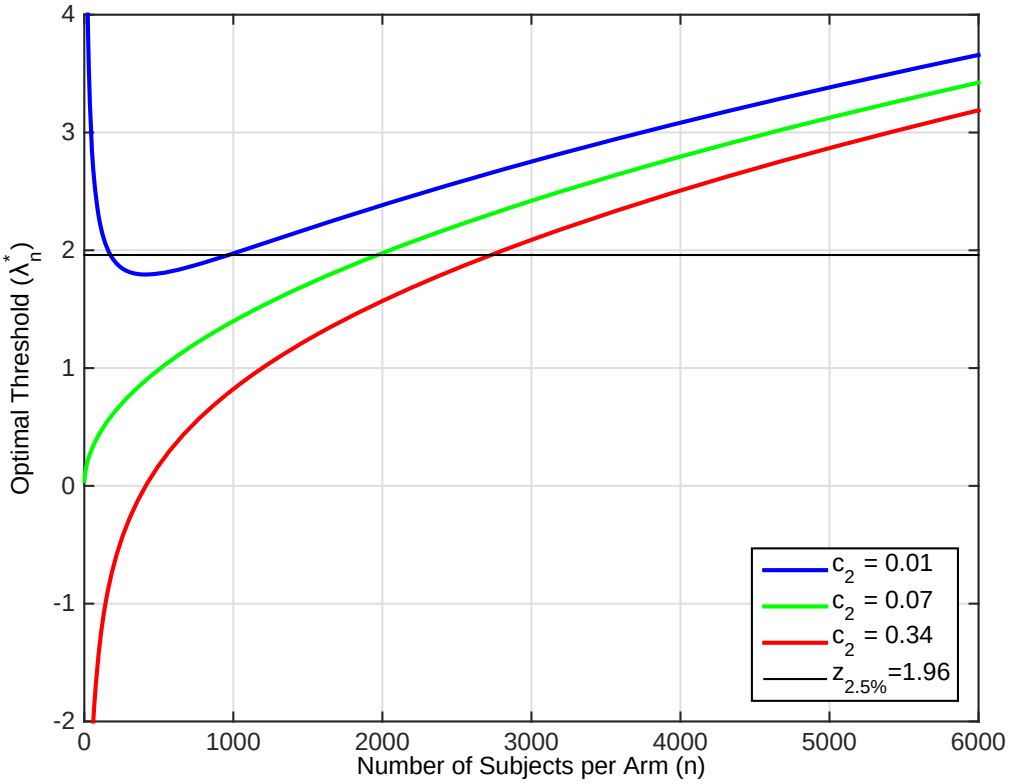


Figure 4: The optimal critical value as a function of the number of subjects per arm for three different diseases. The severity of disease is denoted as c_2 where $c_2 = 0.01$, corresponding to $\overline{c}_2 = 0.2$, represents mild severity. Medium severity corresponds to $c_2 = 0.07$, or equivalently $\overline{c}_2 = 1$, and life-threatening disease is denoted by $c_2 = 0.34$, corresponding to $\overline{c}_2 = 5$. The constant line with the height $z_{2.5\%} = 1.96$ (thin black line) is also drawn for comparison.

Now, if we assume equally likely hypotheses, i.e., $p_0 = p_1 = 0.5$, the parameter $\overline{c}_2$ becomes the ratio of Type II cost to Type I cost, which must be larger for life-threatening diseases than for mild diseases, as discussed in Section 4. In other words, the parameter $\overline{c}_2$ can be considered as a normalized indicator of the severity of the target disease. The more dangerous the disease, the higher the value of $\overline{c}_2$ should be, and the larger chance we should

give to an effective drug to be approved. Therefore, across all sample sizes in Figure 4, by increasing the value of $\overline{c}_2$, the optimal critical value becomes smaller and moves toward the mean of the Z-statistic under the null hypothesis, namely, the constant zero line. In other words, the optimal critical value becomes less conservative as the importance of Type II cost relative to Type I cost increases, modeling a more life-threatening disease. This explains why the red line lies completely below the green line and the green curve is below the blue line. If $\overline{c}_2$ is large enough, the optimal critical value may even cross the zero line and become negative, e.g., if $\overline{c}_2 = 5$, λ_n^* becomes negative over sample sizes smaller than 779. Finally, for $\overline{c}_2 < 1$, implying a larger weight for the Type I cost, corresponding to mild diseases, the behavior of the optimal critical value is qualitatively different from the other two cases, in which the optimal critical value is monotonically increasing in the sample size.

Using the optimal critical value in (8), the size, α, and the power of the test at the alternative hypothesis, $1-\beta$, are given by

$$\alpha = \Phi\left(\frac{1}{\delta_0\sqrt{\mathcal{I}_n}}\log\left(\overline{c}_2\right) - \frac{\delta_0\sqrt{\mathcal{I}_n}}{2}\right), \tag{9}$$

$$1-\beta = \Phi\left(\frac{1}{\delta_0\sqrt{\mathcal{I}_n}}\log\left(\overline{c}_2\right) + \frac{\delta_0\sqrt{\mathcal{I}_n}}{2}\right). \tag{10}$$

Next, for a given n, the expected cost obtained by using the optimal critical value, λ_n^* in (8), can be calculated by substituting (8) into (5) and is given by

$$C(\texttt{fxd}(n,\lambda_n^*)) = p_0 c_1\left[N\Phi(-\lambda_n^*) + N\overline{c}_2\Phi(\lambda_n^* - \delta_0\sqrt{\mathcal{I}_n}) + n(1+\gamma N\overline{c}_2)\right], \tag{11}$$

where the optimal sample size should be determined to minimize this expected cost over all possible sample sizes. Let us consider a continuum of values, rather than discrete values, for the sample size, n, and take the partial derivative of the expected cost in (11) with respect to n as the following:

$$\begin{aligned}\frac{\partial}{\partial n}C(\texttt{fxd}(n,\lambda_n^*)) = &\left(\frac{\partial}{\partial n}\lambda_n^*\right)N\left[-\phi(-\lambda_n^*) + \overline{c}_2\phi(\lambda_n^* - \delta_0\sqrt{\mathcal{I}_n})\right] \\ &- \left(\delta_0\frac{\partial}{\partial n}\sqrt{\mathcal{I}_n}\right)N\overline{c}_2\phi(\lambda_n^* - \delta_0\sqrt{\mathcal{I}_n}) + (1+\gamma N\overline{c}_2),\end{aligned} \tag{12}$$

where the first line is proportional to (7) and, therefore, equal to zero. By simplifying (12)

and setting it to zero to evaluate the optimal sample size n^*, we have:

$$\frac{\partial}{\partial n} C(\texttt{fxd}(n, \lambda_n^*))\Big|_{n=n^*} = \left[-\left(\delta_0 \frac{\partial}{\partial n} \sqrt{\mathcal{I}_n} \right) N\overline{c}_2 \phi(\lambda_n^* - \delta_0 \sqrt{\mathcal{I}_n}) + (1 + \gamma N \overline{c}_2) \right] \Big|_{n=n^*} = 0. \tag{13}$$

Now, if we define $x^* \triangleq \frac{1}{2}\left(\delta_0 \sqrt{\mathcal{I}_{n^*}}\right)^2$, and rearrange terms in (13), then x^* can be represented as the fixed point of a function, g, i.e., $x^* = g(x^*)$. This function is given by

$$g(x) = A \exp\left(-\frac{1}{2} \left(\frac{\log^2(\overline{c}_2)}{x} + x \right) \right), \tag{14}$$

where

$$A = \frac{N^2}{16\pi} \left(\frac{\delta_0^2}{2\sigma^2} \right)^2 \left[\frac{\overline{c}_2}{(1 + \gamma N \overline{c}_2)^2} \right] \tag{15}$$

and $\overline{c}_2 = \frac{p_1 c_2}{p_0 c_1}$ as defined earlier. Now, if N is large enough to make $\gamma N \overline{c}_2$ much larger than 1, A becomes independent of N. Since the exponential function in g is independent of N as well, we observe the insensitivity of the optimal sample size to the prevalence of disease, N, in the case of large burden of disease, i.e., large $C_2 = N c_2$. In the following, we revisit the three cases, for which the optimal critical value is drawn in Figure 4. Let us consider, for all these cases, a target population of $N = 500{,}000$ patients, an alternative hypothesis associated with $\delta_0 = \frac{\sigma}{8}$ and equal prior probabilities for the two hypotheses, i.e., $p_0 = p_1$. We consider $\overline{c}_2 = 0.2$ (equivalently $c_2 = 0.01$) corresponding to an innocuous disease, $\overline{c}_2 = 1$ (equivalently $c_2 = 0.07$) representing a disease with medium severity, and $\overline{c}_2 = 5$ (equivalently $c_2 = 0.34$) corresponding to a life-threatening disease. By using (14), we first determine the optimal sample size, then substitute this n^* into (8) to determine the optimal critical value, and finally, using (9) and (10), we calculate the size of the optimal tests, and their power for the alternative hypothesis. The results are tabulated in Table 3.

In the following section, we employ the cost model proposed in Section 4 and the results of this section to determine the implicit costs in the current standards of clinical trials.

Severity	Optimal Sample Size	Optimal Critical Value	Size (%)	Power (%)
0.01	2,719	2.654	0.40	97.47
0.07	2,236	2.090	1.83	98.17
0.34	1,534	1.266	10.28	98.59

Table 3: The optimal sample size, critical value, size, and statistical power for three trials, each designed to test a treatment targeting a disease with a different severity. For the three trials, the size of the target population is $N = 500{,}000$ and the alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{8}$, for which the power is reported.

A.2 Imputing the Cost of Type I and Type II Errors

We consider a typical one-sided fixed-sample test and assume that it is a BDA-optimal test in our framework using some unknown normalized cost parameter $\overline{c}_2$ and unknown prevalence, N, and infer these parameters for the trial. The FDA regulations require that one-sided tests have at most 2.5% probability of Type I error. For this current standard, it is easy to see that the critical value in a fixed-sample test, on the Z-scale, is

$$\lambda_n^* = z_\alpha \triangleq \Phi^{-1}(1-\alpha), \tag{16}$$

where $\alpha = 2.5\%$ and hence, $z_\alpha = 1.960$. Also, because the Type II error associated with δ_0 is equal to β, we have:

$$\beta = \Phi(\lambda_n^* - \delta_0\sqrt{\mathcal{I}_n}) \quad \Rightarrow \quad z_\beta \triangleq \Phi^{-1}(1-\beta) = \delta_0\sqrt{\mathcal{I}_n} - \lambda_n^* = \delta_0\sqrt{\mathcal{I}_n} - z_\alpha. \tag{17}$$

Substituting (16) and (17) into (8) gives us:

$$z_\alpha = (z_\alpha + z_\beta)^{-1}\log\left(\frac{p_0 c_1}{p_1 c_2}\right) + \frac{z_\alpha + z_\beta}{2} \quad \Rightarrow \quad \log\left(\frac{p_0 c_1}{p_1 c_2}\right) = \frac{z_\alpha^2 - z_\beta^2}{2}. \tag{18}$$

This yields the ratio of Type II cost to Type I cost as:

$$\overline{c}_2 = \exp\left(\frac{z_\beta^2 - z_\alpha^2}{2}\right) = \exp\left(\frac{1}{2}\left(\frac{\delta_0^2}{2\sigma^2}\right)n - z_\alpha\sqrt{\frac{\delta_0^2}{2\sigma^2}n}\right). \tag{19}$$

Note that the cost ratio depends on the number of subjects recruited in the trial. However, in our model, this cost ratio is an exogenous variable which is related to the severity of the targeted disease, the state of current therapies for the disease, and the side effects of the drug. Therefore, the ratio should not depend on the sample size.

Now, in classical hypothesis testing, $\lambda_n^* = z_\alpha$, which is independent of the sample size. Using this fact, we can further simplify the conditions for optimal sample size by noting that the optimal sample size n^*, is the integer value n, for which:

$$\frac{C(\texttt{fxd}(n+1, \lambda_{n+1}^*))}{C(\texttt{fxd}(n, \lambda_n^*))} \geq 1 \quad \text{and} \quad \frac{C(\texttt{fxd}(n-1, \lambda_{n-1}^*))}{C(\texttt{fxd}(n, \lambda_n^*))} > 1. \tag{20}$$

By expanding the left-hand side of (20), we have:

$$\begin{aligned}
\frac{C(\texttt{fxd}(n+1, \lambda_{n+1}^*))}{C(\texttt{fxd}(n, \lambda_n^*))} &= \frac{N\Phi(-z_\alpha) + N\overline{c}_2\Phi\Big(z_\alpha - (z_\alpha + z_\beta)\sqrt{\frac{n+1}{n}}\Big) + (n+1)(1+\gamma N\overline{c}_2)}{N\Phi(-z_\alpha) + N\overline{c}_2\Phi(-z_\beta) + n(1+\gamma N\overline{c}_2)} \\
&= 1 + \frac{(1+\gamma N\overline{c}_2) - N\overline{c}_2\Big[\Phi(-z_\beta) - \Phi\Big(-(z_\alpha+z_\beta)\Big[\sqrt{1+\frac{1}{n}} - 1\Big] - z_\beta\Big)\Big]}{C(\texttt{fxd}(n, \lambda_n^*))} \\
&= 1 + \frac{(1+\gamma N\overline{c}_2) - N\overline{c}_2 \Pr(-z_\beta - \epsilon_1 < Z \leq -z_\beta)}{C(\texttt{fxd}(n, \lambda_n^*))} \geq 1,
\end{aligned} \tag{21}$$

where $Z \sim \mathcal{N}(0,1)$ and $\epsilon_1 = (z_\alpha + z_\beta)\left[\sqrt{1+\frac{1}{n}} - 1\right]$. Furthermore, by expanding the second inequality in (20), we have:

$$\begin{aligned}
\frac{C(\texttt{fxd}(n-1, \lambda_{n-1}^*))}{C(\texttt{fxd}(n, \lambda_n^*))} &= \frac{N\Phi(-z_\alpha) + N\overline{c}_2\Phi\Big(z_\alpha - (z_\alpha + z_\beta)\sqrt{\frac{n-1}{n}}\Big) + (n-1)(1+\gamma N\overline{c}_2)}{N\Phi(-z_\alpha) + N\overline{c}_2\Phi(-z_\beta) + n(1+\gamma N\overline{c}_2)} \\
&= 1 + \frac{N\overline{c}_2\Big[\Phi\Big((z_\alpha+z_\beta)\Big[1-\sqrt{1-\frac{1}{n}}\Big] - z_\beta\Big) - \Phi(-z_\beta)\Big] - (1+\gamma N\overline{c}_2)}{C(\texttt{fxd}(n, \lambda_n^*))} \\
&= 1 + \frac{N\overline{c}_2 \Pr\left(-z_\beta < Z \leq -z_\beta + \epsilon_2\right) - (1+\gamma N\overline{c}_2)}{C(\texttt{fxd}(n, \lambda_n^*))} > 1,
\end{aligned} \tag{22}$$

where $\epsilon_2 = (z_\alpha + z_\beta)\left[1 - \sqrt{1-\frac{1}{n}}\right]$. Next, combining (21) and (22) yields:

$$\Pr(-z_\beta - \epsilon_1 < Z \leq -z_\beta) \leq \frac{1+\gamma N\overline{c}_2}{N\overline{c}_2} < \Pr(-z_\beta < Z \leq -z_\beta + \epsilon_2). \tag{23}$$

Now, both ϵ_1 and ϵ_2 can be well-approximated by $\epsilon_1 \approx \epsilon_2 \approx \frac{z_\alpha + z_\beta}{2n} = \left(\frac{1}{2}\sqrt{\frac{\delta_0^2}{2\sigma^2}}\right)\sqrt{n^{-1}}$. Hence, for a relatively large sample size,[4] n, we can simplify (23) to yield:

$$\frac{1+\gamma N\overline{c}_2}{N\overline{c}_2} \approx \left(\frac{z_\alpha + z_\beta}{2n}\right)\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z_\beta^2\right)\right] = \frac{1}{2\sqrt{2\pi n}}\sqrt{\frac{\delta_0^2}{2\sigma^2}}\exp\left(-\frac{1}{2}z_\beta^2\right). \tag{24}$$

Now, we multiply the result in (24) by (19) to get the following ratio:

$$\frac{1+\gamma N\overline{c}_2}{N} \approx \frac{1}{2\sqrt{2\pi n}}\sqrt{\frac{\delta_0^2}{2\sigma^2}}\exp\left(-\frac{1}{2}z_\alpha^2\right), \quad \text{for large } n. \tag{25}$$

To summarize, for a balanced two-arm fixed-sample test with n subjects per arm and a size of α, i.e., $\texttt{fxd}(n, z_\alpha)$, which has a power of $1-\beta$ at δ_0, we can estimate the normalized Type II cost and prevalence of the disease as:

$$\overline{c}_2 = \exp\left(\frac{z_\beta^2}{2} - \frac{z_\alpha^2}{2}\right) = \exp\left(\frac{1}{2}\left(\frac{\delta_0^2}{2\sigma^2}\right)n - z_\alpha\sqrt{\frac{\delta_0^2}{2\sigma^2}n}\right), \tag{26}$$

$$N \approx \frac{1}{\frac{1}{2\sqrt{2\pi n}}\sqrt{\frac{\delta_0^2}{2\sigma^2}}\exp(-\frac{1}{2}z_\alpha^2) - \gamma\exp(\frac{z_\beta^2 - z_\alpha^2}{2})}, \tag{27}$$

where N is the size of the target population of the drug under test. The expression in (26) is identical to (19), and is repeated for the reader's convenience.

To put the results in (26) and (27) into perspective, let us assume that a fixed-sample test is required with size $\alpha = 2.5\%$, and a power of $1-\beta = 85\%$ for an alternative hypothesis corresponding to $\delta_0 = \frac{\sigma}{8}$. Using the classical hypothesis-testing calculations, this leads to a sample size of n = 1,150 which meets the FDA's criterion. Now let us assume a non-informative prior, i.e., $p_0 = p_1 = 0.5$. For this trial, we get the following severity and prevalence:

$$c_2 = 0.02, \quad N = 15{,}119. \tag{28}$$

Having obtained a severity equal to 0.02 in (28), we can conclude that the current standards for clinical trials are optimal for testing only innocuous diseases, as discussed in Sections 2

[4]For the numerical example given at the end of this section, if the number of recruited patients is more than 100, n is large enough for this approximation to hold.

Power (%)	Required Sample Size	Implied Severity	Implied Prevalence (Thousands)
80	1,005	0.01	13.68
85	1,150	0.02	15.12
90	1,345	0.02	17.51
95	1,664	0.04	24.60

Table 4: The required sample size, implied severity, and prevalence of the target disease for four conventional trials. Each trial corresponds to a different power for the alternative hypothesis, namely, $1-\beta = 80\%, 85\%, 90\%, 95\%$, and all the trials have a size of $\alpha = 2.5\%$. For all the trials, the alternative hypothesis corresponds to $\delta_0 = \frac{\sigma}{8}$.

and 4, and cannot be optimal for more life-threatening diseases like pancreatic cancer. In general, as observed in Table 3, for a given target population, the test should become less conservative (its critical value should become smaller) and the sample size should shrink as the severity of the disease increases to avoid exposure to inferior treatment during the trial. Now, maintaining all our assumptions and only changing the power of the test for the alternative hypothesis, we get different values for the required sample size, implied severity, c_2, and prevalence, N. We have reported the results for four different power levels for the alternative hypothesis, namely, $1-\beta = 80\%, 85\%, 90\%$, and 95%, in Table 4. As observed in Table 4, all the implied severity values for these classical tests are too small, especially, for a high power level of 95% where the implied severity is only 0.04 (last row). These small numbers underscore the fact that the current standards of clinical trials are quite conservative and not suitable for terminal illnesses with no effective treatment.