Patient-Centered Clinical Trial Design for Heart Failure Devices via Bayesian Decision Analysis

Patient-Centered Clinical Trial Design for Heart Failure Devices via Bayesian Decision Analysis

Running Heading: Patient-Centered Clinical Trial Design

Shomesh E. Chaudhuri, PhD[1], Phillip Adamson, MD[2], Dean Bruhn-Ding, BS[3], Zied Ben Chaouch, PhD[4,5], David Gebben, PhD[6], Liliana Rincon-Gonzalez, PhD[7], Barry Liden, JD[8], Shelby D. Reed, PhD[9,10], Anindita Saha, BS[6], Daniel Schaber, PharmD[11], Kenneth Stein, MD[12], Michelle E. Tarver, MD, PhD[6], Andrew W. Lo, PhD[1,4,13–15]

[1]QLS Advisors, Cambridge, MA
[2]Abbott Laboratories, Abbott Park, IL
[3]CVRx, Inc., Minneapolis, MN
[4]MIT Laboratory for Financial Engineering, Cambridge, MA
[5]MIT Department of Electrical Engineering and Computer Science, Cambridge, MA
[6]FDA Center for Devices and Radiological Health, Silver Spring, MD
[7]Medical Device Innovation Consortium, Arlington, VA
[8]Edwards Lifesciences, Irvine, CA
[9]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC
[10]Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC
[11]Medtronic, Minneapolis, MN
[12]Boston Scientific, Marlborough, MA
[13]MIT Sloan School of Management, Cambridge, MA
[14]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA
[15]Santa Fe Institute, Santa Fe, NM

Address for correspondence:
Andrew W. Lo
245 Main Street
Cambridge, MA 02142
alo@qlsadvisors.com

Total word count: 4,420 (max: 6,000 words)

Abstract (337 words, max: 250 to 450 words)

Background: The statistical significance of clinical trial outcomes is generally interpreted quantitatively according to the same threshold of 2.5% (in one-sided tests) to control the false positive rate or type I error, regardless of the burden of disease or patient preferences. The clinical significance of trial outcomes—including patient preferences—are also considered, but through qualitative means that may sometimes be challenging to reconcile with the statistical evidence.

Objective: To apply Bayesian decision analysis (BDA) to heart failure device studies to choose an optimal significance threshold that maximizes the expected utility to patients across both the null and alternative hypotheses, thereby allowing clinical significance to be incorporated into statistical decisions either in the trial design stage or in the post-trial interpretation stage. In this context, utility is a measure of how much well-being the approval decision for the treatment provides to the patient.

Methods: We use the results from a discrete-choice experiment (DCE) study focusing on heart failure patients' preferences, questioning respondents about their willingness to accept therapeutic risks in exchange for quantifiable benefits with alternative hypothetical medical device performance characteristics. These benefit-risk tradeoff data allow us to estimate the loss in utility—from the patient perspective—of a false positive or false negative pivotal trial result. We compute the BDA-optimal statistical significance threshold that maximizes the expected utility to heart failure patients for a hypothetical two-arm, fixed-sample, randomized control trial (RCT). An interactive Excel-based tool is provided that illustrates how the optimal statistical significance threshold changes as a function of patients' preferences for varying rates of false positives and false negatives, and as a function of assumed key parameters.

Results: In our baseline analysis, the BDA-optimal significance threshold for a hypothetical two-arm RCT with a fixed sample size of 600 patients per arm was 3.2%, with a statistical power of 83.2%. This result reflects the willingness of heart failure patients to bear additional risks of the investigational device in exchange for its probable benefits. However, for increased device-associated risks and for risk-averse subclasses of heart failure patients, BDA-optimal significance thresholds may be smaller than 2.5%.

Conclusions: Bayesian decision analysis is a systematic, transparent, and repeatable process for combining clinical and statistical significance, explicitly incorporating burden of disease and patient preferences into the regulatory decision-making process.

Key words: Bayesian decision analysis; patient preferences; heart failure; medical devices; benefit-risk.

Key Points for Decision Makers

- Bayesian decision analysis is applied to heart failure device studies to select an optimal type I error threshold that maximizes the patient's expected utility and which is consistent with the patient population's risk preferences.
- In our baseline analysis for a hypothetical two-arm RCT of a heart failure device with a fixed sample size of 600 patients per arm, we find a BDA-optimal significance threshold of 3.2%, which is above the 2.5% commonly used threshold, and a statistical power of 83.2%.
- The BDA approach explicitly incorporates burden of disease and patient preferences into the regulatory decision-making process, and BDA-optimal significance thresholds may be used directly to inform statistical decisions either in the trial design stage or in the post-trial interpretation stage.

Abbreviations List: BDA, bayesian decision analysis; CDRH, Center for Devices and Radiological Health; CRT, cardiac-resynchronization therapy; DCE, discrete-choice experiment; FDA, US Food and Drug Administration; HF, heart failure; ICD, implantable cardioverter-defibrillator; LVADs, left ventricular assist devices; MDIC, Medical Device Innovation Consortium; NYHA, New York Heart Association; RCT, randomized control trial.

Declarations

I. Funding

Funding was coordinated by the Medical Device Innovation Consortium (MDIC) with contributions from Abbott, Abiomed, Boston Scientific, CVRx, Inc., Edwards Lifesciences, Medtronic, and the U.S. Food and Drug Administration via subcontract between Quantitative Life Sciences Advisors and MDIC as part of MDIC's Science of Patient Input in Medical Device Clinical Trials.

II. Conflicts of Interest

SEC is a co-founder and principal of QLS Advisors LLC, a healthcare investments advisor, and QLS Technologies LLC, a healthcare analytics and consulting company

AWL reports personal investments in private biotechnology companies, biotechnology venture capital funds, and mutual funds. AWL is a co-founder and principal of QLS Advisors LLC, a

healthcare investments advisor, and QLS Technologies LLC, a healthcare analytics and consulting company; a director of AbCellera, Annual Reviews, Atomwise, BridgeBio Pharma, and Roivant Sciences; and an advisor to Apricity Health, Aracari Bio, BrightEdge Impact Fund, Enable Medicine, FINRA, Lazard, NIH/NCATS, Quantile Health, SalioGen Therapeutics, Swiss Finance Institute, and Thalēs. During the most recent six-year period, AWL has received speaking/consulting fees from AbCellera, AlphaSimplex Group, Annual Reviews, Apricity Health, Aracari Bio, Atomwise, Bernstein Fabozzi Jacobs Levy Award, BridgeBio, Cambridge Associates, Chicago Mercantile Exchange, Enable Medicine, Financial Times Prize, Harvard Kennedy School, IMF, Journal of Investment Management, Lazard, National Bank of Belgium, New Frontier Advisors/Markowitz Award, Oppenheimer, Princeton University Press, Q Group, QLS Advisors, Quantile Health, Research Affiliates, Roivant Sciences, SalioGen Therapeutics, Swiss Finance Institute, and WW Norton.

The other authors have no conflict of interest to disclose aside from their employment with the sponsoring companies.

III. Availability of Data and Material

To help readers better understand how Bayesian decision analysis works, we provide an easy-to-use BDA optimization tool in Excel. This tool allows users to recompute the results using their own parameter values of interest.

IV. Ethics Approval

The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate.

V. Consent to Participate

NA.

VI. Consent for Publication

NA.

VII. Code Availability

To help readers better understand how Bayesian decision analysis works, we provide an easy-to-use BDA optimization tool in Excel. This tool allows users to recompute the results using their own parameter values of interest.

VIII. Author Contributions

Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work: all authors. Drafting the work or revising it critically for important intellectual content: all authors. Final approval of the version to be published: all authors. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: all authors.

Acknowledgements

We thank the Medical Device Innovation Consortium (MDIC), sponsor companies (Abbott Laboratories, Abiomed, Boston Scientific, CVRx, Edwards Lifesciences, and Medtronic), the Duke Clinical Research Institute, and the US Food and Drug Administration (FDA) Center for Devices and Radiological Health (CDRH) for collecting and providing access to patient-reported data, and Jayna Cummings for editorial support.

1. Background

The regulatory process for the market authorization of medical diagnostic and therapeutic products relies on a framework where benefits and risks are weighed against one another. The consequences of approving an ineffective or unsafe product (a "type I error," or a false positive) with potentially dangerous side-effects must be weighed against not approving a safe and effective product (a "type II error," or a false negative) that could help treat or diagnose the disease for patients. Regulators must strike the proper balance between these two types of errors. They do so by considering multiple factors, including quantitative information such as valid, scientific evidence of safety and effectiveness. However, they also consider other information such as the burden of disease on the population, the risk of adverse events post-approval, the current standard of clinical care and available alternatives, and the extent to which the patient is willing to trade off the potential benefits and risks. How regulators weigh these considerations to render a regulatory decision may foster criticism by stakeholder groups when they disagree with the outcome. This process can be made more systematic and transparent by applying Bayesian decision analysis (BDA) to incorporate qualitative considerations into the traditional quantitative analysis of clinical trial results, as described in a series of recent publications (1–6, 20, 21). In particular, the BDA methodology has been applied in collaboration with regulatory agencies to weight loss devices (3), Parkinson's Disease for deep brain stimulation devices (6), and wearable dialysis devices for dialysis-dependent kidney disease (21). The detailed preference studies used in (6) and (21) were analyzed through interval regressions. From a methodological perspective, (5) proposes an extended formulation to the BDA which accounts for the patient's risk aversion, but also for the patient's uncertainty aversion (which can be induced by the uncertainty in the device's attributes).

To describe how BDA incorporates these considerations, it is helpful to contrast BDA with the traditional quantitative approach for conducting a statistical test of the null hypothesis of no effectiveness: we choose a desired type I error rate, typically 2.5% for one-sided tests (or 5% for two-sided tests) (17), and evaluate the statistical significance of the clinical evidence against this threshold. If the results are inconsistent with the null hypothesis at a significance level or p-value less than 2.5%, then the null hypothesis is rejected, and in this context, the product may be authorized. This is, of course, an oversimplification of the regulatory process, done for expositional convenience. In practice, regulators use many contextual factors such as clinical significance in addition to statistical significance when deciding whether to authorize a medical device. The

clinical significance here is defined by whether the stakeholders of the study find the results meaningful. A clinically significant intervention is one where the effects are large enough to justify the associated costs, inconveniences, and harms worthwhile (18, 19). While statistical significance may be a necessary condition of clinical significance, it is not a sufficient condition.

The point of this example is that choice of 2.5% is purely arbitrary and is mostly used for historical reasons (17) and by itself does not reflect any of the qualitative considerations such as the severity of the disease to be treated by the medical device. BDA aims to determine a type I error rate that is appropriate for a specific treatment for a specific disease. In the case of life-limiting diseases with no or poor existing treatment options, patients and other stakeholders may be willing to accept a higher false positive rate when testing a new treatment, especially if it yields a lower false negative rate. Conversely, for less severe diseases, the decrease in patient utility associated with approving an ineffective and possibly harmful treatment could be considerably larger than rejecting an effective one resulting in lower maximum accepted false positive rate. In the BDA framework, the statistical threshold is determined by explicitly minimizing the expected loss in utility to patients due to both type I and type II errors, where the expected loss in utility is the sum of the measured impact of false positives (e.g., the adverse effects experienced by patients exposed to an ineffective and potentially harmful product) and false negatives (e.g., the disease burden of patients who could have benefited from the product), each weighted by their respective probabilities.

BDA requires more information than the traditional approach: the losses in utility under both types of error must be specified, and these losses may be difficult to gauge. However, several theoretically sound approaches can be used for this purpose, including stated-preference surveys designed to quantify patients' preferences. This inclusion of additional information into the process is how qualitative considerations can be more formally quantified, allowing regulators to incorporate burden of disease, risk assessments, and patient preferences into their decision-making process in an explicit, transparent, and systematic way.

In this analysis, we illustrate this approach using the results from a discrete-choice experiment (DCE) that quantified heart failure (HF) patients' willingness to accept medical device risks in exchange for improved probable benefits of medical devices (7). In the supplemental materials, we provide an interactive Excel-based tool that demonstrates how the optimal statistical

significance threshold changes as a function of the utility lost from approving ineffective and possibly harmful product and the failure to approve a safe and effective product, as determined by the HF patient preference information.

2. Methods

For this clinical trial design, we consider a quantitative utility-based framework that takes patient preferences into explicit account across multiple device attributes when determining the optimal statistical significance threshold of a balanced two-arm fixed-sample Randomized Controlled Trial (RCT). Although we have assumed a balanced trial for expositional simplicity, this methodology can also be applied more generally to single-arm trials using objective performance criteria from prior data, as well as multi-armed unbalanced trials with certain modifications.

We first define a patient-centered utility model (utility defined as satisfaction or well-being resulting from an approval decision on the device) associated with given medical device attributes, including benefits and device-associated risks and features. Like Chaudhuri et al. (3, 6), our patient utility model is based on preference data for a specific but hypothetical device. This model, detailed in Appendix A, is based on a previously described framework applied to oncology trials (1, 2), and can be used in other contexts in which patient preference data are available.

We assign prior probabilities to each combination of attribute levels under both the null and alternative hypotheses and formulate the expected utility of the trial under the assumption that the device would be approved if and only if the trial provides sufficient statistical evidence of efficacy. The optimal one-sided significance level (α, or critical value λ_α) is then determined to maximize the expected utility of the trial. Maximizing the utility of the trial means concluding either that the device's probable benefits outweigh the risks for the proposed indications for use, or that the probable benefits do not outweigh the risks. Incidentally, maximizing the expected utility of the trial is equivalent to minimizing the trial's expected loss of utility, which includes the consequences of incorrect decisions for current and future patients.

2.1 Patient Utility Model

A multidisciplinary team[1] developed a survey administered to patients with HF to estimate their utility score when presented with distinct hypothetical medical devices with different medical attributes. One of the attributes in the study was the number of additional years with the current limitation of physical activity due to HF symptoms equivalent to severity Class II or III of the New York Heart Association (NYHA) Functional Classification system (see https://heart-failure.net/symptoms-stages). The DCE survey was administered to quantify the impact of an increase in the rate of mortality or device complications on the patient's overall utility, as well as the change in utility due to a one-year increase in survival at the same physical functional level (i.e., corresponding to marked physical activity limitations consistent with an NYHA III classification) and/or a one-year increase in survival with improved physical functional level (i.e., corresponding to moderate physical activity limitations consistent with an NYHA II classification).

The patient preference weights depicted in **Figure 1**, estimated from the survey data, can be aggregated across device attributes to compute an overall utility score for a specific treatment. For example, **Figure 1** shows that the patient's decrease in utility due to an increase in the rate of device-associated mortality and/or complications for a more invasive surgery (which can be interpreted as risks) could be offset by additional years in NYHA II or III according to the patient preference information (which can be interpreted as benefits). The relative loss of utility per patient, L, of using a lower-scored intervention over another, is then defined in terms of the net difference in the patient's utility (see Appendix A).

[1] The team includes researchers from the Medical Device Innovation Consortium (MDIC), sponsor companies (Abbott Laboratories, Abiomed, Boston Scientific, CVRx, Edwards Lifesciences, and Medtronic), the Duke Clinical Research Institute, and the US Food and Drug Administration (FDA) Center for Devices and Radiological Health (CDRH).

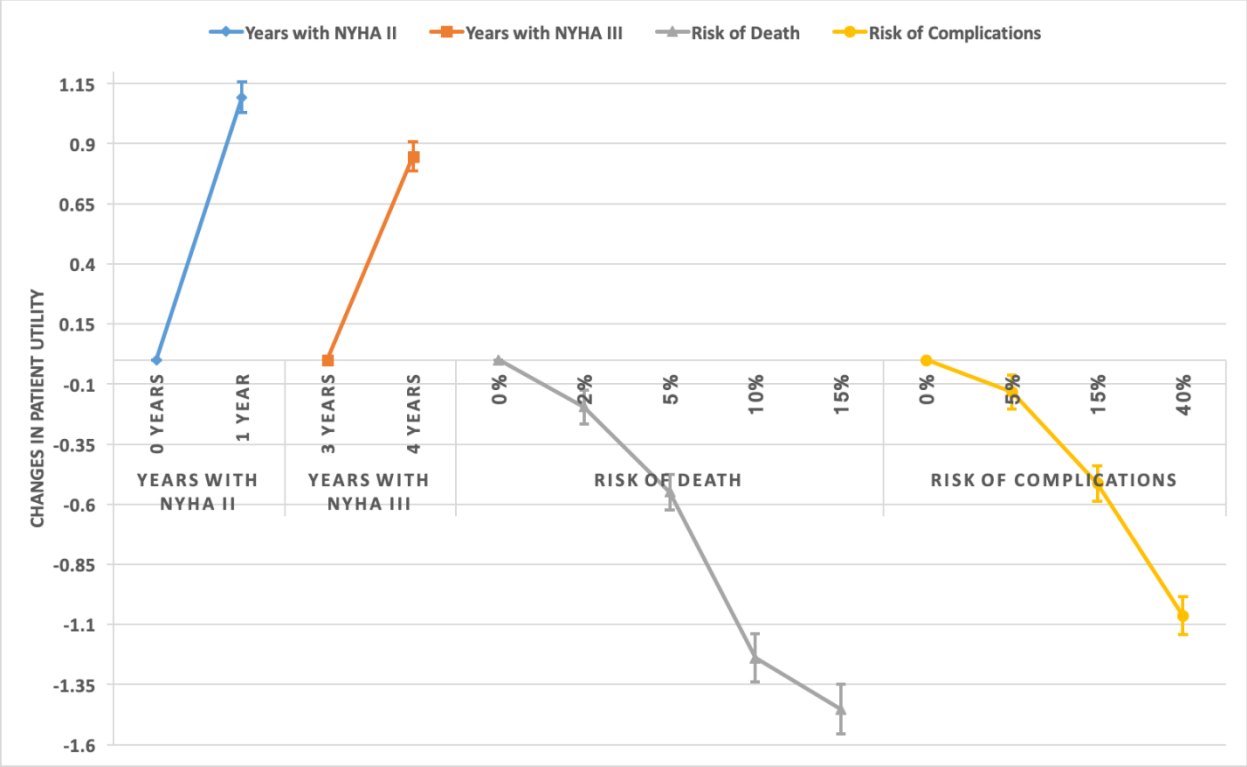

Figure 1. Patient preference model estimated for heart failure patients (7). A DCE technique was used to quantify patients' changes in utility as we vary each device attribute (holding the other attributes constant). These device attributes include increases in the rate of complications or death due to the device, as well as one-year increases in survival at the same and/or at an improved physical functional level (according to the New York Heart Association Functional Classification system). **Abbreviations List**: NYHA, New York Heart Association.

The BDA framework uses these utility scores to estimate the utility lost from the patient's perspective of making an incorrect approval decision. For example, in the case of an incorrect approval, the new device is assumed to provide no increase in survival or in NYHA functional Class II or III relative to the standard of care but does provide an increase in the rate of device associated mortality and/or complications, as well as potential missed opportunities to be treated by therapies with a higher utility score. We denote by L_1 the utility lost per patient as the result of an incorrect approval, i.e., a "false positive." On the other hand, a failure to authorize a safe and effective device (i.e., a "false negative") results in the missed opportunity to gain from a therapy that is associated with a higher utility score than that of the standard of care. This loss is denoted L_2. A detailed description of the utility-based framework is included in Appendix A. The potential

loss of utility per patient of an incorrect decision (i.e., either approving an unsafe or ineffective device, or the failure to authorize a safe and effective device) is shown in **Table 1**. Although regulators use additional contextual factors to inform their decision, we make the simplifying assumption that the regulator would approve a device only if its effectiveness is shown to be statistically significant. The utility severity ratio, L_1/L_2, provides a measure of their relative importance. Multiplying these losses by their probabilities and summing across the various scenarios results in the expected loss of utility of a potentially incorrect approval decision. The number of patients affected by the incorrect decision can be used to scale these values to estimate a collective loss of utility. In this case study, we assume that the size of the patient populations affected by a type I or type II error are approximately equal, hence our focus on the per patient loss in utility. The BDA framework determines the optimal statistical significance threshold such that the expected loss in utility of these downside scenarios is minimized.

	Do Not Approve	Approve
Device Equally or Less Effective than the Control (H_0)	0	L_1 (Type I Error)
Device More Effective than the Control (H_1)	L_2 (Type II Error)	0

Table 1. Estimated loss of utility per patient associated with a clinical trial. We assume there is no post-trial loss of utility for a correct decision, i.e., not approving (approving) a device that is equally (more) effective than the control.

2.2 *Bayesian Decision Analysis*

The BDA framework is applied to the primary effectiveness endpoint for the trial; however, the same framework can be applied to evaluate safety endpoints such as an increase in the rate of device-associated 30-day mortality and complications. Similar to Tang et al (8), we assume the primary effectiveness endpoint is reduction in the rate of death from any cause or hospitalization for HF over an observation period of 40 months. We further assume that subjects in the treatment arm receive the investigational device, and each subject's response is independent of all other responses. In the same way, patients in the control arm are assumed to receive standard of care treatment.

Assuming an exponential distribution for the time to event (i.e., death from any cause or hospitalization for HF) for each patient given a particular treatment, we have the following expression for the mean of the log-rank statistic in the Cox proportional hazard regression under the alternative hypothesis (δ_n),

$$\delta_n = -\frac{1}{2}\log(r)\sqrt{\sum_{k=0}^{1}\sum_{i=0}^{n-1} d_{i,k}}\,, \qquad [1]$$

where r denotes the hazard ratio, and $d_{i,k}$ is the probability that a subject in trial arm k will suffer an event during the observation period. Under the alternative hypothesis, subjects in the control arm ($k = 0$) have a higher event rate than subjects in the experimental arm ($k = 1$) who receive the investigational device. Therefore, $d_{i,0} = 1 - \exp(-h_0 o_{i,0})$ and $d_{i,1} = 1 - \exp(-h_1 o_{i,1})$, where h_k and $o_{i,k}$ are the event rate and observation period for subject i in trial arm k, respectively. Under the assumption that the number of observed events is sufficiently large, the log-rank statistic is approximately normal. The log-rank statistic, Z, is then compared to the critical value, λ_α. Finding that $Z > \lambda_\alpha$ supports the rejection of the null hypothesis.

Assuming previously observed probabilities p_0 and p_1 (where $p_0 + p_1 = 1$) for the cases where the investigational device is less or equally effective ($H = 0$) or more effective ($H = 1$) than the control treatment, and letting V_0 and V_1 be the utility in the hypothetically optimal scenarios where the correct approval decision is made, it is straightforward to calculate the expected utility associated with an RCT design with parameters (n, λ_α) as

$$\begin{aligned}\mathrm{E}[\text{Utility}; n, \lambda_\alpha] = {} & p_0(\mathrm{V}_0 - \mathrm{E}[\text{Loss in utility} \mid H = 0]) \\ & + p_1(V_1 - \mathrm{E}[\text{Loss in utility} \mid H = 1])\end{aligned} \qquad [2]$$

where

$$\mathrm{E}[\text{Loss in utility} \mid H = 0] = \alpha \cdot L_1, \quad [3]$$

$$\mathrm{E}[\text{Loss in utility} \mid H = 1] = \beta \cdot L_2, \quad [4]$$

α is the significance level, and $(1 - \beta)$ is the statistical power of the trial (i.e., this measures the ability of the study to correctly rejecting the null hypothesis when the alternative hypothesis is true). The optimal critical value (λ_α^*) is determined such that the expected utility of the trial is maximized. Finally, in solving the optimization problem, we observe that the expected utility of the trial is maximized when the expected loss in utility, $\mathrm{E}[\text{Loss in utility}; n, \lambda_\alpha] = p_0\ \mathrm{E}[\text{Loss in utility} \mid H = 0] + p_1\ \mathrm{E}[\text{Loss in utility} \mid H = 1]$, is minimized.

2.3 Assumptions of the BDA Framework

We summarize in **Table 2** the parameter values used in our analysis. These parameters have been chosen to align with literature reviews of effectiveness and safety of HF devices (8-11).

First, we assume that the investigational device is either non-effective (i.e., the null hypothesis) or effective (i.e., the alternative hypothesis) with equal prior probability ($p_0 = p_1 = 50\%$). This is consistent with the principle of clinical equipoise, which states that there is genuine uncertainty in the expert medical community over whether a treatment will be effective.

Next, we assume that, if effective, the device will provide an extra year of functional equivalence to NYHA III compared to the control treatment, but with a 0.5% and 10.0% increase in the rate of device-associated 30-day mortality and complications, respectively. The calibration of these parameters relies on quantitative and qualitative input from scientists and physicians with domain-specific expertise.

Finally, like Tang et al. (8), who randomly assigned patients with NYHA class II or III HF to receive either an implantable cardioverter-defibrillator (ICD) alone, or an ICD plus cardiac-resynchronization therapy (CRT) we assume a primary effectiveness endpoint of a reduction in either the rate of death from any cause or hospitalization for HF and an observation period of 40 months. We further assume the annualized event rate of the primary effectiveness endpoint is 40.3% in the control arm, and 33.2% in the investigational arm (8). The target accrual is set to

1,200 subjects (both arms) to achieve a baseline statistical power of 80% given a one-sided alpha value of 2.5%.

We conduct sensitivity analyses to evaluate the robustness of our analysis to perturbations of the key parameter values assumed by our model. To help readers better understand how Bayesian decision analysis works, we provide an easy-to-use BDA optimization tool in Excel[2]. This tool allows users to recompute the results using their own parameter values of interest.

Parameter	Description	Value
Probability that the device is more effective than the control (p_1)	The estimated *a priori* probability that the device is more effective than the control ($H = 1$), which can be estimated from historical success rates or set to 50% when there is no prior information	50%
NYHA II duration increase (in years)	Additional duration of functioning equivalent to NYHA II under $H = 1$ compared to the control group	0.0
NYHA III duration increase (in years)	Additional duration of functioning equivalent to NYHA III under $H = 1$ compared to the control group	1.0
Rate of mortality	Increase in the rate of device-associated 30-day mortality	0.5%
Rate of complications	Increase in the rate of a collection of potential complications that could occur to account for various potential adverse events	10.0%
Control group event-rate (h_0)	Annualized event rate of primary effectiveness endpoint (death from any cause or hospitalization for heart failure) in the control group	40.3%
Treatment group event-rate (h_1)	Annualized event rate of primary effectiveness endpoint (death from any cause or hospitalization for heart failure) in the treatment group	33.2%
Observation Period (o) (in months)	Observation follow-up time for primary endpoint	40
Target accrual ($2n$)	Total number of patients in both arms of the trial	1,200

Table 2. Assumptions for heart failure device RCT design. Abbreviations List: NYHA, New York Heart Association.

[2] Tested on Microsoft Excel Version 2205 Build 16.0.15225.20172 64-bit

3 Results

Using BDA and the estimated patient utility model, we are able to formulate patient-centered fixed-sample RCTs for HF devices.

3.1 *Baseline Results*

Figure 2 depicts the relation between type I error and the expected loss in utility to patients (Panel A), and the relation between type I error and power (Panel B), as well as the specific values of these variables for the standard RCT and the BDA-optimal decision rule. In particular, Panel A in **Figure 2** reports that the BDA-optimal one-sided significance threshold for the hypothetical trial is 3.2%, which is 28% greater than the threshold value of 2.5% for the standard design. This result reflects the implicit willingness of these patients to bear additional uncertainty regarding the effectiveness of the investigational device so as to reduce the chance of missing out on its probable benefits. Given a fixed sample size of 600 patients in each arm of the trial, this results in a clinical trial with a statistical power of 83.2% (see Panel B in **Figure 2**).

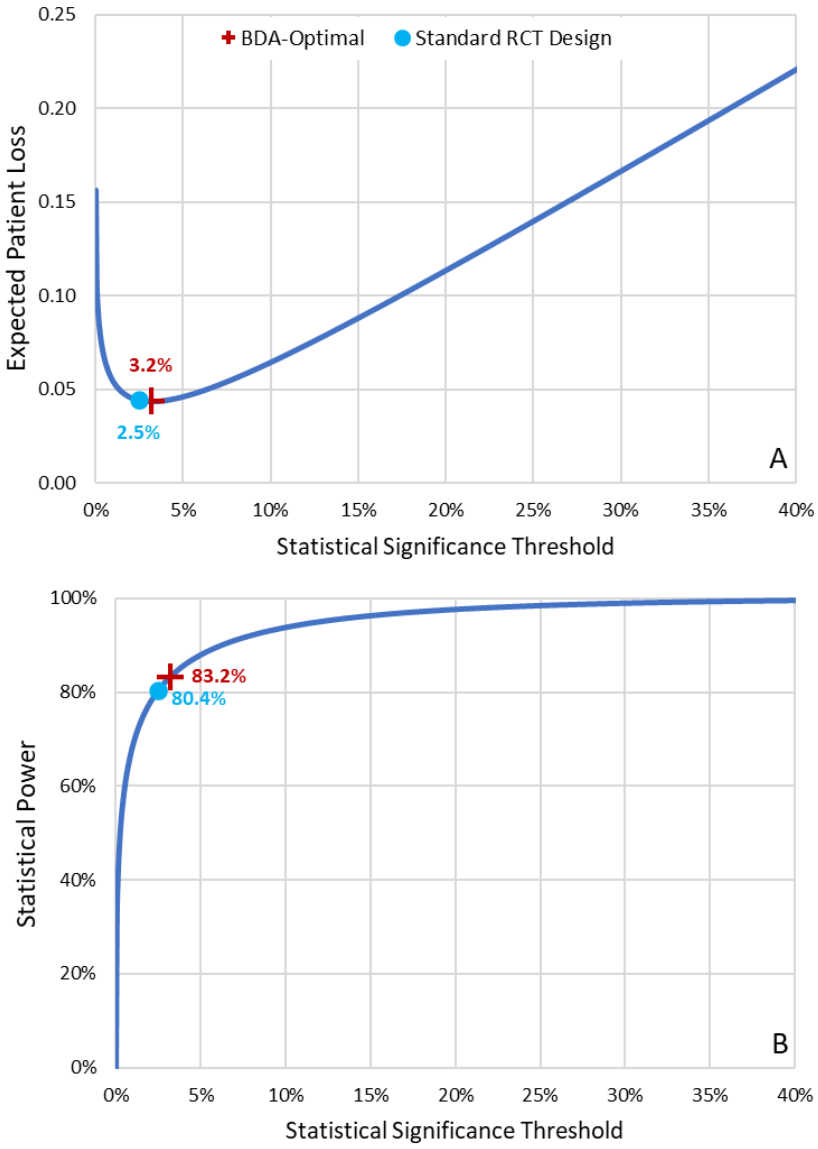

Figure 2. Standard and BDA-optimal RCT for a heart failure device. Relation between type I error and the expected loss in utility to patients (Panel A), between type I error and power (Panel B). The specific values of these variables for the standard RCT and the BDA-optimal decision rule are highlighted in blue and red respectively. The BDA-optimal one-sided significance threshold for the hypothetical trial is 3.2%, which is 28% greater than the threshold value of 2.5% for the standard design, and a statistical power of 83.2% (given a fixed sample size of 600 patients in each arm of the trial). **Abbreviations List**: BDA, bayesian decision analysis; RCT, randomized control trial.

3.2 Sensitivity Analysis

In this section, we investigate the robustness of our results to the parameter assumptions of our model. As we vary the estimated attributes of the investigational device, we update the BDA-

optimal trial design. The optimal significance level and power of the perturbed parameters are reported in **Table 3**, where the sensitivity analysis corresponds to deviations from the baseline rate originally reported in **Table 2**. For example, as the increase in device-associated mortality rate varies from 0% to 1%, which is approximately the rate for most HF devices with the exception of left ventricular assist devices (LVADs), the BDA-optimal significance threshold varies from a maximum of 4.5% to a minimum of 2.0%. This result shows the relative robustness of the baseline BDA-optimal significance threshold of 3.2% to changes in this key input for this device because as the risk increases, patients' tolerance to potential lack of benefit decreases, as expected.

Finally, the BDA framework can also be used to identify the BDA-optimal significance threshold for patient groups with different preferences. For the more risk-tolerant group identified in latent-class analysis in the preference study, which among other characteristics tend to include patients that previously received a cardiac device (7), we find that the BDA-optimal one-sided significance threshold for our hypothetical trial is 9.4%, which is almost 4 times the standard threshold of 2.5%. For patients with prior experience with cardiac devices, utility is lost due to clinical trials that are too conservative about their false approval rate.

Conversely, for risk-averse patients, more often those with no previous cardiac device experience, the traditional significance level of 2.5% is more permissive than the calculated patient-centered thresholds. In fact, given the baseline parameters in **Table 3**, the expected utility under both the null and alternative hypotheses is negative for these patients. Hence, the BDA framework would recommend this investigational device be rejected without conducting a trial. In this scenario, patients would require additional utility under the alternative hypothesis before considering the device as a possible treatment.

	Severity Ratio (L_1/L_2)	**BDA-Optimal Significance**	**BDA-Optimal Power**
Increase in NYHA II duration (years)			
0.00	3.52	3.2%	83.2%
0.10	2.23	4.6%	87.0%
0.20	1.63	5.7%	89.2%
0.30	1.29	6.7%	90.6%

Increase in NYHA III duration (years)			
0.70	—	—	—
0.80	36.26	0.4%	54.7%
0.90	6.41	2.0%	77.4%
1.00	3.52	3.2%	83.2%
Increase in mortality rate from device			
0.00%	2.29	4.5%	86.8%
0.25%	2.81	3.8%	85.1%
0.50%	3.52	3.2%	83.2%
0.75%	4.55	2.6%	80.9%
1.00%	6.19	2.0%	77.7%
Increase in the rate of complications			
5.0%	1.05	7.7%	91.8%
7.5%	1.82	5.3%	88.4%
10.0%	3.52	3.2%	83.2%
12.5%	10.30	1.3%	71.7%
15.0%	—	—	—

Table 3. Sensitivity of the BDA-optimal one-sided significance threshold and statistical power to perturbations of the attributes of the investigational device. Abbreviations List: BDA, bayesian decision analysis; NYHA, New York Heart Association.

4 Discussion

In this study, we found that a fixed significance level of 2.5% in a trial does not necessarily maximize patient utility. Under base case assumptions, the BDA-optimal one-sided significance threshold for the hypothetical trial was 3.2%. This result reflects the willingness of HF patients to bear additional risk of the investigational device to reduce the chance of missing out on its probable benefits. For these patients, utility is potentially lost due to trials that are too conservative about their type I error rate. Conversely, for patient subclasses with more risk-aversion, traditional thresholds of 2.5% may be too permissive. In these cases, patients would require a clear demonstration of clinical effectiveness to reduce the probability of approving an ineffective device and subsequent harm to their health.

Two practical issues in applying BDA must be addressed in any specific application: calibrating BDA input parameters to the expected losses in utility and addressing the consequences

of a larger number of false positives. The former can be addressed by convening "calibration advisory panels," composed of representatives from several stakeholder communities including patient advocacy groups, to provide input to regulators. Incorporating patient preferences in the regulatory approval process may also be important for medical devices that require significant commitment on the part of the patient to ensure adherence to specific treatment regimens.

For example, if risk-tolerant HF patients were willing to bear the risk of a novel therapeutic technology under development, this preference could be factored into the regulatory review process. Surveys of HF patients, such as the one used in this analysis, in which the participants were asked to choose between the current standard of care versus other hypothetical treatments, could help to determine the strength of patient preferences, which could then serve as one of several inputs for informing the meaningfulness of the statistical significance of the trial results during the regulatory decision process.

The ability of the BDA framework to systematically weigh multifaceted tradeoffs that reflect a variety of perspectives combined with its flexibility and practicality make it a potentially valuable tool for both study design and post-trial analysis. In this article, we highlight the post-trial application of BDA, where the study design is based on current best practices and where we use BDA-optimal significance thresholds as a tool to help summarize the totality of the data—including patient preferences—at the decision-making stage. This mitigates risk as the study would remain valid even if the BDA framework was not ultimately applied. Alternatively, clarity around calibrating the loss in utility of false positives and negatives may help device manufacturers design their trials to incorporate these considerations most efficiently. In this case, the BDA tool could be used during the design stage with a fixed power, and jointly optimized over the sample size and alpha level of the study. This would require that all decision rules are determined *a priori*, preferably with regulatory guidance.

One barrier we foresee to the uptake of BDA is the lack of existing, high-quality, relevant PPI data, which may not available to a researcher in the early development stages. It may be beneficial to include the acquisition of PPI within a clinical trial plan or in a parallel protocol using an external sample and have the analysis plan prospectively incorporated into the clinical trial statistical analysis plan (SAP) to support clinical assumptions used to design the trial.

5 Limitations

Our findings must be qualified in several respects. First, for expositional simplicity, we have only considered traditional two-arm, fixed-sample RCT. In practice, trial designs may include multiple arms and adaptive features involving interim analysis and adjustments for early signals of effectiveness, futility, or adverse effects. These modifications may alter the optimal type I and II error rates and appropriate modifications to our calculations are required to determine the BDA-optimal designs for these settings (5).

Second, while we have made strong assumptions in this case study for illustrative purposes, these assumptions can be readily relaxed or modified in future applications. For example, when considering potential regulatory decisions for the broader population, aggregating the preferences of patient subgroups by prevalence, incidence rates, and other epidemiological measures within this framework could be considered. In addition, patient utility models that incorporate diminishing marginal returns, present-biased time preferences, and other factors can also be incorporated into the BDA model (1–6). We believe that a nuanced consideration of these issues will be instructive in the design of future clinical trials.

Third, the trials considered here use a primary effectiveness endpoint of a reduction in either the rate of death from any cause or hospitalization for HF, both of which are of clear and unambiguous clinical importance. In cases where combined endpoints have variable importance to patients, or a large gradient in frequency of occurrence, definitions of the severity of type I or type II error and assumptions about how these endpoints correlate with other device attributes require further consideration.

Finally, we have confined our attention to statistical significance thresholds. As shown in its benefit-risk guidance, the FDA/CDRH currently considers a variety of factors beyond p-values when making its decisions (13 – 15). While the analysis conducted here under the assumption that assumption that the regulator would approve a device only if its effectiveness is shown to be statistically significant, it is important to note that regulators such as the FDA and the EMA have approved drugs or medical devices despite mixed statistical results by taking other factors into account such as burden of disease and patient preferences. As an illustration, we can refer to the accelerated approval of aducanumab (Aduhelm [Biogen]), a drug for Alzheimer's disease, in June 2021 despite debates on its statistical effectiveness and the significance of certain clinical endpoints. With very limited alterative treatments available for Alzheimer’s disease, patient

advocacy groups have been advocating for the drug's approval. Another controversy arose in 2016 following the approval of eteplirsen (Exondys 51 [Sarepta Therapeutics]), a drug for Duchenne muscular dystrophy. This progressively debilitating disease is a rare genetic disorder characterized by progressive muscle degeneration and weakness, typically affecting young boys and with an average life expectancy of around 25-30 years. As for aducanumab, at the time, there had been no drug approvals for Alzheimer's disease since 2003.

However, determining the maximum acceptable level of uncertainty for clinical evidence in a systematic way remains an unresolved question in regulatory science for all stakeholders. As such, the ability of BDA to systematically and transparently weigh tradeoffs that reflect a multiplicity of perspectives, and from that, to calculate the optimal threshold for the significance level, makes this framework a potentially valuable tool for facilitating patient-centered clinical trial designs. The FDA/CDRH encourages sponsors to discuss their novel clinical trial plans early through its Q-Submission Program (16).

6 Conclusions

This study indicated that, for the population surveyed in (7), the traditional statistical significance threshold of 2.5% did not necessarily maximize patient utility. Respondents were willing to trade higher therapeutic risks for greater benefit. This could indicate that in the studied population, traditional trial designs may be too conservative and overly focused on avoiding false-positive results. This focus on avoiding false-positive results can result in missed opportunities to approve safe and effective medical devices. Conversely, for more risk-averse patient groups, the current thresholds of just statistical significance may be more permissive than BDA-optimal thresholds. The relative size of those populations must be taken into account for appropriately applying the BDA thresholds to the selected population under evaluation.

While the BDA framework is robust, we emphasize that careful consideration must be applied to the assumptions underlying the specific models in order to produce useful recommendations. If correctly implemented, the BDA perspective has the potential to benefit all stakeholders.

References

1. Isakov L, Lo AW, Montazerhodjat V. Is the FDA too conservative or too aggressive?: A Bayesian decision analysis of clinical trial design. *Journal of Econometrics*, 211(1), pp. 117-136. 2019.

2. Montazerhodjat V, Chaudhuri SE, Sargent DJ, Lo AW. Use of Bayesian decision analysis to minimize harm in patient-centered randomized clinical trials in oncology. *JAMA Oncology*, 3(9). 2017.

3. Chaudhuri SE, Ho MP, Irony T, Sheldon M, Lo AW. Patient-centered clinical trials. *Drug Discov Today*, 02, 23(2), pp. 395-401. 2017

4. Chaudhuri, S. E., Lo, A. W., Xiao, D., & Xu Q. Bayesian Adaptive Clinical Trials for Anti-Infective Therapeutics during Epidemic Outbreak. *Harvard Data Science Review* [Internet]. 2020;2. Available from: https://doi.org/10.1162/99608f92.7656c213

5. Ben Chaouch, Z., Chaudhuri, S. E. and Lo, A.W. Bayesian Decision Analysis under Risk and Uncertainty: a Tale of Two Exposures. Manuscript under review. 2022.

6. Chaudhuri, S. E., Ben Chaouch, Z., Hauber, B., Mange, B., Zhou, M., Ho, M., Saha, A., Caldwell, B., Benz, H. L., Ruiz, J., Christopher, S., Bardot, D., Sheehan, M., Donnelly, A., McLaughlin, L., Gwinn, K., Sheldon, M. & AWL. Use of Bayesian decision analysis to maximize value in patient-centered randomized clinical trials in Parkinson's Disease. *Journal of Biopharmaceutical Statistics*. 2022.

7. Reed, S. D., Yang, J. C., Rickert, T., Johnson, F. R., Gonzalez, J. M., Mentz, R. J., Krucoff, M. W., Vemulapalli, S., Adamson, P. B., Gebben, D. J., Rincon-Gonzalez, L., Saha, A., Schaber, D., Stein, K. M., Tarver, M. E., & Bruhn-Ding, D. Quantifying Benefit-Risk Preferences for Heart Failure Devices: A Stated-Preference Study. *Circulation Heart failure*, 15(1):e008797. 2022.

8. Tang, A. S., Wells, G. A., Talajic, M., Arnold, M. O., Sheldon, R., Connolly, S., Hohnloser, S. H., Nichol, G., Birnie, D. H., Sapp, J. L., Yee, R., Healey, J. S., Rouleau, J. L., & Resynchronization-Defibrillation for Ambulatory Heart Failure Trial Investigators. Cardiac-resynchronization therapy for mild-to-moderate heart failure. *The New England Journal of Medicine*, 363(25), 2385–2395. 2010.

9. Stone, G. W., Lindenfeld, J., Abraham, W. T., Kar, S., Lim, D. S., Mishell, J. M., Whisenant, B., Grayburn, P. A., Rinaldi, M., Kapadia, S. R., Rajagopal, V., Sarembock, I. J., Brieke, A., Marx, S. O., Cohen, D. J., Weissman, N. J., Mack, M. J., & COAPT Investigators. Transcatheter Mitral-Valve Repair in Patients with Heart Failure. *The New England Journal of Medicine*, 379(24), 2307–2318. 2018.

10. Abraham, W. T., Kuck, K. H., Goldsmith, R. L., Lindenfeld, J., Reddy, V. Y., Carson, P. E., Mann, D. L., Saville, B., Parise, H., Chan, R., Wiegn, P., Hastings, J. L., Kaplan, A. J., Edelmann, F., Luthje, L., Kahwash, R., Tomassoni, G. F., Gutterman, D. D., Stagg, A., Burkhoff, D., Hasenfuß, G. A Randomized Controlled Trial to Evaluate the Safety and Efficacy of Cardiac Contractility Modulation. *JACC Heart Failure*, 6(10), 874–883. 2018.

11. Zile, M. R., Lindenfeld, J., Weaver, F. A., Zannad, F., Galle, E., Rogers, T., & Abraham, W. T. Baroreflex Activation Therapy Patients with Heart Failure with Reduced Ejection Fraction. *Journal of the American College of Cardiology*, 76(1), 1–13. 2020.

12. Lo, A.W. Discussion: New directions for the FDA in the 21st century. *Biostatistics*, 06, 18(3), pp. 404-407. 2017.

13. FDA. Factors to Consider When Making Benefit–risk Determinations in Medical Device Premarket Approval and De Novo Classifications. 2016. Available from: https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm506679.pdf

14. FDA. Patient Preference Information — Voluntary Submission, Review in Premarket Approval Applications, Humanitarian Device Exemption Applications, and De Novo Requests, and Inclusion in Decision Summaries and Device Labeling. 2016. Available from: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM446680.pdf

15. FDA. Consideration of Uncertainty in Making Benefit-Risk Determinations in Medical Device Premarket Approvals, De Novo Classifications, and Humanitarian Device Exemptions – Guidance for Industry and Food and Drug Administration Staff. August 2019 https://www.fda.gov/regulatory-information/search-fda-guidance-documents/consideration-uncertainty-making-benefit-risk-determinations-medical-device-premarket-approvals-de

16. FDA. Requests for Feedback and Meetings for Medical Device Submissions: The Q-Submission Program—Guidance for Industry and Food and Drug Administration Staff. 2019. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/requests-feedback-and-meetings-medical-device-submissions-q-submission-program

17. Fisher, R. A. *The design of experiments.* Oliver & Boyd. 1935 (6th edition: 1953).

18. Ranganathan, P., Pramesh, C. S., & Buyse, M. Common pitfalls in statistical analysis: Clinical versus statistical significance. *Perspectives in clinical research*, vol. 6,3, pp.169-70. doi:10.4103/2229-3485.159943. 2015.

19. Ferreira, M. L. and Herbert, R. D. What does 'clinically important' really mean?. *Australian Journal of Physiotherapy*, 54.4, pp. 229-230. 2008.

20. Xu, Q., Cho, J., Ben Chaouch, Z. & Lo, A.W.: Incorporating patient preferences and burden-of-disease in evaluating ALS drug candidate AMX0035: a Bayesian decision analysis perspective, *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 2022. DOI: 10.1080/21678421.2022.2136994

21. Ben Chaouch, Z., Xu, Q., Chaudhuri, S.E., Gebben, D. J., Harris, R.C., Flythe, J.E., Hurst, F.P., Mansfield, C., Saha, A., Sheldon, M., Siah, K.W., Tarver, M.E., Treiman, K., West, M., Wood, D., and Lo, A.W.: Use of Bayesian Decision Analysis in the Design of Patient-Centered Clinical Trials for Kidney Failure Devices. Manuscript under review. 2023