For Informational Purposes Only

Patient-Centered Clinical Trial Design Tool for Heart Failure Devices

REPORT PREPARED FOR MDIC

Produced By

Quantitative Life Sciences

www.qlsadvisors.com

Table of Contents

Overview

The regulatory process for the market authorization of medical diagnostic and therapeutic products is fraught with ethical dilemmas not faced by regulators outside the healthcare industry. The consequences of approving an ineffective therapy with potentially dangerous side-effects (a "type I error," or a false positive) must be weighed against not approving a safe and effective therapy (a "type II error," or a false negative) that could help ease the burden of disease for many patients. Regulators must strike the proper balance between these two types of errors by considering multiple factors, including scientific merit, the clinical evidence from trials, especially randomized control trials (RCTs) of the therapy under review, the burden of disease, the current standard of care and its alternatives, and patient preferences. Given the complexity of biomedicine and the potential consequences of both types of error, regulators must exercise a certain degree of discretion and flexibility when making their decisions. But how these factors are—and should be—weighed is not always clear, a process which only encourages criticism by stakeholder groups that might disagree with the decision.

However, this process can be made more transparent and systematic by applying Bayesian decision analysis (BDA) to regulatory approval decisions, as described in a series of recent publications (1)(2)(3)(4)(5)(6).

The simplest description of BDA involves comparing it to the simplest version of the traditional approach for conducting a statistical test of the null hypothesis of no effectiveness: choose a desired type I error rate, say 2.5%, 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 our context, the therapy is approved.

The question raised and addressed by BDA is "why 2.5%?" For fatal diseases that have no existing treatment or poor treatments, 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, as is often the case. In the BDA framework, the regulatory approval threshold is determined by explicitly minimizing the expected loss to patients due to both type I and type II errors, where the expected loss is the sum of the measured impact of false positives (e.g., the side effects experienced by patients exposed to an ineffective therapy) and false negatives (e.g., the disease burden of patients who could have benefited from the therapy), each weighted by their respective probabilities.

BDA does require more information than the traditional approach: the losses under both types of error must be specified, and in some cases, these losses may be difficult to gauge. However, several metrics have been developed for just this purpose, including survey tools designed by patient advocacy groups to measure the preferences of their members.

In this analysis, we apply the results from a survey of heart failure patients that quantifies the maximum level of risk they would be willing to accept in order to achieve different potential therapeutic benefits (7). The interactive QLS BDA Tool illustrates how the optimal statistical significance threshold changes as a function of the cost of false approval and false rejection, as determined by the heart failure patient preference information.

Methodology

For this clinical trial design, we consider a quantitative 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 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 value model associated with given medical device attributes. Like Chaudhuri et al. (3)(6), our patient value model is based on preference data for a specific but hypothetical device. This framework is derived from previous applications to oncology trials (1)(2), and can be used in other contexts in which patient preference data is available.

We assign prior probabilities to each possible combination of attributes, and formulate the expected value of the trial. The optimal one-sided significance level (α or critical value, λ_α) is then determined to maximize the expected value of the trial. Note that here, maximizing the value of the trial means either providing access to a safe and effective treatment to patients, or concluding that the treatment has not demonstrated a reasonable assurance of safety and effectiveness. This is equivalent to minimizing potential losses of the trial, which include the consequences of incorrect decisions for all patients.

Patient Value Model

A multidisciplinary team, including researchers from the Medical Devices 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) developed and administered a survey to patients with heart failure to quantify the maximum level of risk that patients would accept in order to achieve different potential benefits of a heart failure device (7). The aim of this study was to apply best-practice stated preference methods to quantify the willingness of heart failure patients to accept therapeutic risks in exchange for improved effectiveness. Based on this input from patients and regulators, a discrete choice experiment (DCE) survey was developed and administered to quantify the benefit-risk tradeoffs that are acceptable to people with heart failure. The DCE technique was used to quantify patients' risk tolerance

for complications or death in exchange for increases in functional years in NYHA II or NYHA III heart failure. Figure 1 shows the patient preference model estimated from the survey data.

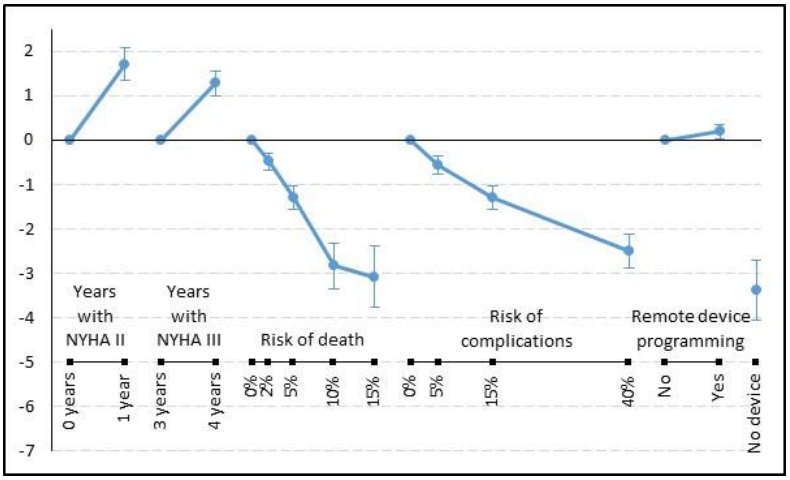

Figure 1. Patient preference model estimated for heart failure patients (7). A DCE technique was used to quantify patients' risk tolerance for complications or death in exchange for increases in functional years in NYHA II or NYHA III.

The patient value model estimates the additional risk (in the form of device complications or death) that a patient would be willing to accept before becoming indifferent to receiving a given benefit. The absolute level of benefit includes additional years in NYHA II or NYHA III heart failure. Finally, latent-class analysis was used to specify the model for different subgroups of patients.

The patient preferences depicted in Figure 1 can be aggregated across benefits and risks to compute an overall preference score for a specific treatment. For example, the patient value lost from an increase in mortality risk resulting from a more invasive surgery could be offset by additional benefits. These benefits, such as the value gained from increase in functional years in NYHA III, can be mapped to a patient preference score using Figure 1. Given both these changes, and holding all other attributes constant, the net change in value can then be calculated to determine if the additional years in NYHA III more than compensates for the increased mortality risk according to the patient preference information. The relative loss of value per patient, L, of using a lower-scored intervention over another, is then defined in terms of the net difference in value.

The BDA framework uses these benefit-risk preferences to estimate the value 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 benefits relative to the standard of care, but does provide additional risk in the form of mortality, complications, and missed opportunities to be treated by more effective therapies. We denote this risk L_1,

the value lost per patient as the result of an incorrect approval, i.e., a "false positive". On the other hand, an incorrect failure to approve a device (i.e., a "false negative") results in the missed opportunity to benefit from a therapy that is more effective than the standard of care. This loss is denoted L_2. The potential loss per patient of an incorrect decision is shown in Table 1. The severity ratio, L_1/L_2, provides a measure of their relative importance.

Table 1. Estimated loss in value per patient associated with a clinical trial. We assume there is no post-trial loss in value for a correct decision, i.e., rejecting (approving) a device that is less (more) preferred relative to the control.

	Rejected	Approved
Ineffective	0	L_1
Safe and Effective	L_2	0

Multiplying these costs by their probabilities, and summing across the various scenarios, results in the expected loss of an incorrect regulatory decision. The number of patients affected by the incorrect decision can be used to scale these values to estimate a collective loss of value. In this case study, we assume that these populations are approximately equal, hence our focus on the per patient loss. The BDA framework determines the optimal statistical significance threshold such that the expected harm of these downside scenarios is minimized.

Bayesian Decision Analysis

A quantitative primary endpoint that tests for effectiveness is assumed for the trial; however, the same BDA framework can be applied to evaluate safety concerns such as the risk of device-associated 30-day mortality and complications. Similar to Tang et al (8), we assume the primary endpoint is death from any cause, or hospitalization for heart failure 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 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),

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

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 equally effective ($H = 0$) or more effective ($H = 1$) to the control treatment, and letting V_0 and V_1 be the value created in the hypothetically optimal scenarios where the correct approval decision is made, it is straightforward to calculate the expected value associated with an RCT design with parameters (n, λ_α) as

$$\mathrm{E}[\text{Value}; n, \lambda_\alpha] = p_0(\mathrm{V}_0 - \mathrm{E}[\text{Loss} \mid H = 0]) + p_1(V_1 - \mathrm{E}[\text{Loss} \mid H = 1]) \tag{2}$$

where

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

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

α is the significance level, and $1 - \beta$ is the power of the trial. The optimal critical value (λ_α^*) is determined such that the expected value of the trial is maximized. Finally, in solving the optimization problem, we observe that the expected value of the trial is maximized when the expected loss, $\mathrm{E}[\text{Loss}; n, \lambda_\alpha] = p_0\mathrm{E}[\text{Loss} \mid H = 0] + p_1\mathrm{E}[\text{Loss} \mid H = 1]$, is minimized.

Heart Failure Device Case Study

Using BDA and the estimated patient preference model, we are able to formulate patient-centered fixed-sample RCTs for heart failure devices. Table 2 summarizes the parameter values used in our analysis.

These parameters have been calibrated based on regulatory reviews of heart failure devices and literature reviews of effectiveness and safety (8-11). First, we assume that the investigational device is either ineffective (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 beneficial.

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 an increased 0.5% risk of device-associated 30-day mortality and a 10.0% risk of complications. 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 heart failure to receive either an implantable cardioverter-defibrillator (ICD) alone, or an ICD plus cardiac-resynchronization therapy (CRT), we assume a primary effectiveness endpoint of either death from any cause or hospitalization for heart failure, and an observation period of 40 months. We further assume the annualized event rate of 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 provide readers with greater transparency and intuition behind our Bayesian decision model, we provide an easy-to-use BDA optimization tool in Excel, allowing users to recompute the results using their own parameter values of interest.

Table 2. Assumptions for heart failure device RCT design.

Parameter	Description	Value
Probability that the treatment is effective (p_1)	The estimated *a priori* probability that the treatment is effective ($H = 1$), which can be estimated from historical success rates or set to 50% when there is no prior information	50%
NYHA II benefit	Additional duration (in years) of functioning equivalent to NYHA II under $H = 1$ compared to the control group	0.0
NYHA III benefit	Additional duration (in years) of functioning equivalent to NYHA III under $H = 1$ compared to the control group	1.0
Risk of death	Risk of device-associated 30-day mortality	0.5%
Risk of complications	Risk 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)	Observation follow-up time (in months) for primary endpoint	40
Target accrual ($2n$)	Total number of patients in both arms of the trial	1,200

Baseline Results

Panel A in Figure 1 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 the standard design of 2.5%. This result reflects the willingness of these patients to bear additional uncertainty regarding the effectiveness of the investigational device (i.e., the additional duration of functional equivalence to NYHA III) in order to reduce the chance of missing its potential 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 1).

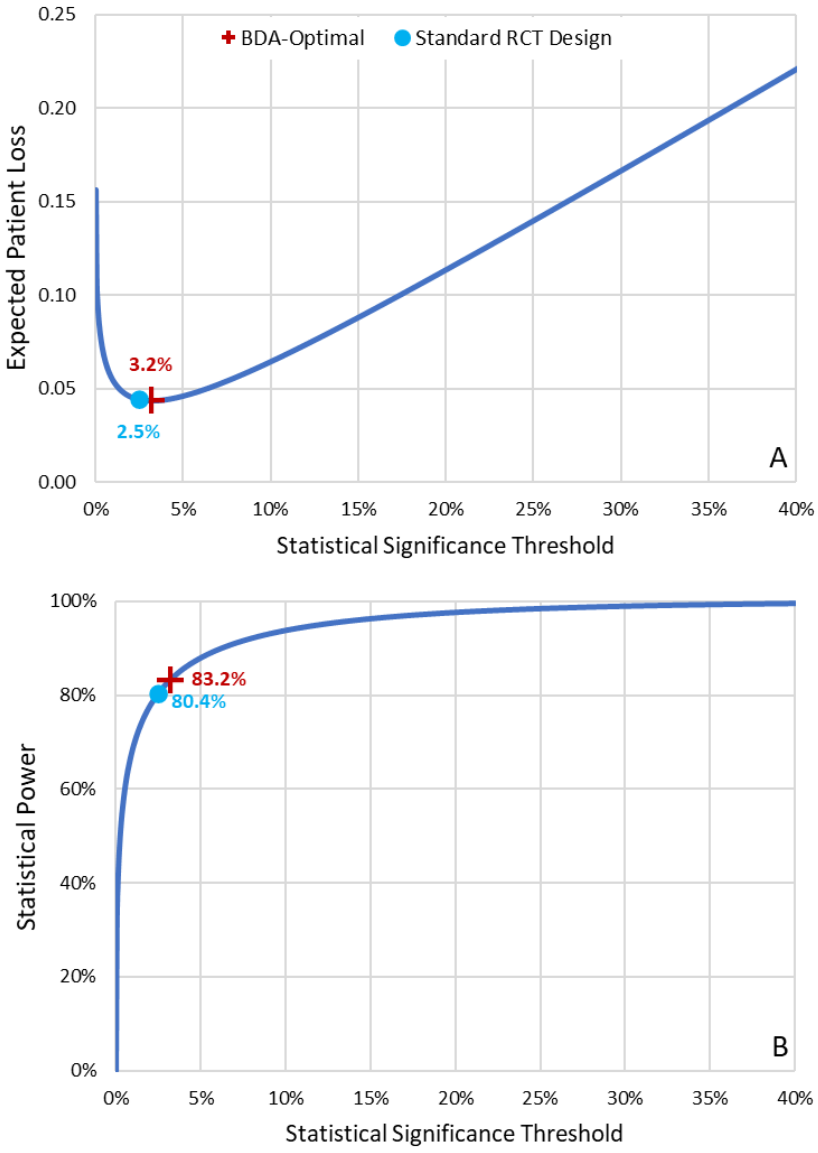

Figure 1. Standard and BDA-optimal RCT for a heart failure device.

Sensitivity Analysis

In this section, we investigate the robustness of our results to the parameter assumptions of our model. As we vary the estimated effectiveness and safety profile 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 risk of device-associated death 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 varies from a maximum of 4.5% to a minimum of 2.0%. This result shows the relative insensitivity of the baseline BDA-optimal significance threshold of 3.2% to changes in this key input for this device.

For the NYHA II benefit, the four-number summary in Table 3 corresponds to [0 years, 0.1 years, 0.2 years, 0.3 years] below the baseline rate of 1.0 years reported in Table 2. We find that, if the investigational device has only a small amount of benefit relative to the control treatment, the investigational device is not preferred to the control treatment under the alternative hypothesis, and the BDA framework recommends the investigational device be rejected without conducting a trial.

Table 3. Sensitivity of the BDA-optimal one-sided significance threshold and statistical power to perturbations of the safety and effectiveness profile of the investigational device.

	Severity Ratio (L_1/L_2)	BDA-Optimal Significance	BDA-Optimal Power
NYHA II benefit (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%
NYHA III benefit (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%
Risk of death			
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%
Risk 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%	—	—	—

Finally, the BDA framework can also be used to identify the BDA-optimal significance threshold for different subgroups of patients. Using the patient preferences for the more risk-tolerant subgroup of patients, 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, value is lost due to clinical trials that are too conservative about their false approval rate. Here, a missed opportunity to approve a potentially beneficial yet more risky treatment has a substantial negative impact on risk-tolerant patients.

Conversely, for risk-averse patients, such as those with no previous cardiac device experience, the traditional significance level of 2.5% is more permissive than the calculated patient-centered thresholds. Indeed, given the baseline parameters in Table 3, the investigational device is not preferred to the control treatment even under the alternative hypothesis. In these cases, patients require clear demonstration of clinical effectiveness to reduce the probability of a false approval and subsequent harm to their health.

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, it is possible to aggregate the preferences of patient subgroups by prevalence, incidence rates, and other epidemiological measures within this framework. In addition, patient value models that incorporate diminishing marginal returns or present-biased time preferences can also be incorporated into the BDA model. A more detailed discussion of these and other factors can be found in (1–6). We believe that a nuanced consideration of these issues will be instructive in the design of future clinical trials.

Discussion

Two practical issues in applying BDA must be addressed in any specific application: calibrating the expected losses, 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, to provide input to regulators, including patient advocacy groups, which have the expertise, direct access to patients, and motivation to accurately represent their members' preferences. Incorporating patient preferences in the regulatory approval process is especially important for medical devices that require significant commitment on the part of the user to ensure adherence to specific treatment regimens.

For example, if risk-tolerant heart failure patients with previous cardiac device experience were willing to bear the risk of a novel therapeutic technology under development, this preference should be factored into the regulatory review process. Surveys of heart failure patients, such as the one used in this analysis, in which the subjects are asked to choose between the current standard of care versus other hypothetical treatments, can help to determine the strength of patient preferences, which can then serve as one of several inputs for determining the optimal regulatory decision.

The issue of higher false positives can be addressed by creating a temporary approval to market a safe and effective therapy with a higher degree of uncertainty that expires after a short period (say, four years) (12). During this period, the sponsor would be required to collect and share data on the performance of its therapy. If the therapy meets or exceeded expectations, the temporary approval converts to a standard approval, otherwise it expires, and the therapy is withdrawn. Of course, the regulator should have the right to terminate the temporary approval at any time in response to adverse events or significantly negative data.

Both calibration advisory panels and short-term temporary approvals would greatly accelerate the pace of therapeutic development for a number of underserved medical needs, including alternative heart failure devices, and neither would limit regulatory flexibility in any way. Regulatory flexibility is particularly important because discretion, judgment, and experience are essential to regulators in deciding which therapies to approve.

Given this flexibility, BDA may seem redundant and unnecessary. However, regulators may still benefit from using a systematic, rational, transparent, repeatable, and practical framework in which their decisions can be clearly understood by and communicated to all stakeholders, while explicitly incorporating feedback from these communities.

References

1. Isakov L, Lo AW, Montazerhodjat V. Is the FDA too conservative or too aggressive?: A Bayesian decision analysis of clinical trial design. J Econom. 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 Oncol. 2017;3(9).

3. Chaudhuri SE, Ho MP, Irony T, Sheldon M, Lo AW. Patient-centered clinical trials. Drug Discov Today. 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 Sci Rev [Internet]. 2020;2. Available from: https://doi.org/10.1162/99608f92.7656c213

5. Chaudhuri, Shomesh E., Lo AW. Bayesian adaptive patient-centered clinical trials. 2019.

6. Chaudhuri, S. E., 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. Under review. 2020.

7. Reed, S. D., et al. Quantifying Benefit-Risk Preferences for Heart Failure Devices: A Stated-Preference Study. Under review. 2020.

8. Tang, A. S. L., et al. Cardiac-resynchronization therapy for mild-to-moderate heart failure. New England Journal of Medicine. 2010.

9. Stone, G.W., et al. Transcatheter Mitral-Valve Repair in Patients with Heart Failure. New England Journal of Medicine. 2018.

10. Abraham, W.T., et al. A Randomized Controlled Trial to Evaluate the Safety and Efficacy of Cardiac Contractility Modulation. 2018.

11. Zile, M.R., et al. Baroreflex Activation Therapy Patients with Heart Failure with Reduced Ejection Fraction. Journal of the American College of Cardiology. 2020.

12. Lo AW. Discussion: New directions for the FDA in the 21st century. Biostatistics. 2017.