

Biostatistical Design in Device Clinical Trials

Marco Muller¹, Matteo Klein², Oscar Hansen³

¹ Professor, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: marco.muller788@yahoo.com | ORCID: 4377-6945-0275-1299

² Research Scientist, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: matteo.klein703@yahoo.com | ORCID: 9710-3294-6237-5896

³ Postdoctoral Researcher, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: oscar.hansen979@yahoo.com | ORCID: 2243-5458-7956-0969

ABSTRACT

Medical device clinical trials present unique biostatistical challenges absent from pharmaceutical trials: operator learning curves confound early performance data, device iteration during development creates mid-trial design changes, patient-device interaction variability exceeds drug pharmacokinetic variability, and objective performance criteria (OPC) based on historical controls replace randomised comparisons for many device categories. Traditional fixed-sample parallel-group designs adapted from drug trials often produce underpowered or inefficient device evaluations because they fail to account for these device-specific complexities. This study presents the Biostatistical Design Assessment Framework for Device Trials (BDAFDT), evaluating five trial design strategies -- fixed-sample single-arm OPC, fixed-sample randomised controlled trial, Bayesian adaptive design with informative priors, group sequential design with interim analyses, and platform trial with master protocol -- across four design efficiency tasks: sample size optimisation, learning curve accommodation, design change flexibility, and evidence strength for regulatory submission. Our Trial Design Efficiency Score (TDES) integrates statistical power, sample efficiency, adaptive capability, regulatory acceptance, and operational feasibility. Bayesian adaptive design achieved the highest TDES (0.926) through informative priors incorporating predicate device data that reduced required sample sizes by 34% while maintaining 90% power, and platform trials achieved the highest design flexibility (0.955) through master protocols enabling seamless addition of new device arms without restarting the trial infrastructure.

Keywords: clinical trial design; medical device; Bayesian adaptive; sample size; learning curve; objective performance criteria; platform trial; group sequential; regulatory evidence; biostatistics

Citation: Muller et al. [2023]. Biostatistical Design in Device Clinical Trials. DOI: <https://doi.org/10.5281/zenodo.19610267>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Jun 19 Accepted: 2023 Aug 01 Published: 2023 Aug 17

Research Article: Research Article

1. Introduction

1.1 Device-Specific Trial Design Challenges

Medical device clinical trials operate under constraints that fundamentally differ from pharmaceutical trials. The operator learning curve effect means that device performance improves systematically as surgeons or interventionalists gain experience with a new device -- the first 20-50 procedures typically show significantly worse outcomes than subsequent procedures, creating a confound that masks the device's true performance if early cases are included in the primary analysis (Vahl et al., 2017). Device iteration during development means that the device evaluated in later trial phases may differ materially from the version tested in early phases -- a challenge absent from drug trials where the molecular entity is fixed. Blinding is often impossible because the device, its implantation technique, and its imaging appearance are recognisable to both patients and clinicians. Comparator selection is complicated by the absence of a placebo-equivalent for surgical devices and the ethical difficulty of sham surgery controls (Sedrakyan et al., 2014). Finally, the FDA's acceptance of single-arm trials with objective performance criteria (OPC) for many device categories -- comparing device outcomes against a pre-specified performance threshold derived from historical data rather than a concurrent control group -- creates statistical design questions about threshold derivation, Type I error control, and generalisability that randomised trial designs do not encounter (Campbell, 2017).

1.2 Five Trial Design Strategies

Five design strategies address these challenges with different statistical frameworks. Fixed-sample single-arm OPC designs enrol a pre-specified number of patients and compare the observed event rate against a performance goal derived from literature or registry data -- the most common design for PMA and 510(k) pivotal trials, used in approximately 65% of FDA device approvals (Dhruva et al., 2009). Fixed-sample randomised controlled trials (RCTs) provide the strongest evidence but require larger samples and longer timelines than single-arm designs. Bayesian adaptive designs use informative prior distributions incorporating predicate device data and allow sample size re-estimation, treatment arm dropping, or early stopping based on accumulating evidence (Berry et al., 2010). Group sequential designs pre-specify interim analyses

with spending functions that control overall Type I error while enabling early stopping for efficacy or futility (Jennison and Turnbull, 2000). Platform trials with master protocols create standing trial infrastructure within which multiple device arms can enter, be evaluated, and exit while sharing a common control arm and operational framework (Woodcock and LaVange, 2017). Section 2 reviews the literature. Section 3 describes BDAFDT. Results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 Evidence Standards for Device Approval

The clinical evidence bar for medical device approval has been a persistent topic of regulatory debate. Dhruva et al. (2009) analysed 78 PMA-approved cardiovascular devices and found that the median pivotal trial enrolled 265 patients with 24 months follow-up, and only 27% were randomised -- substantially less rigorous than pharmaceutical Phase III standards. Rathi et al. (2015) followed up on PMA-approved devices and found that 36% subsequently received safety communications, label changes, or post-market study requirements, raising questions about whether the pre-market evidence base was adequate. The counterargument, articulated by Campbell (2017), is that device trials serve fundamentally different purposes than drug trials: devices are iteratively improved rather than molecularly fixed, device performance depends on operator skill that improves with experience, and the relevant clinical question is often whether the device meets an acceptable performance threshold rather than whether it outperforms a specific comparator. This perspective supports OPC-based designs for many device categories while acknowledging that RCTs remain essential when comparative effectiveness against alternative treatments drives clinical adoption decisions.

2.2 Bayesian and Adaptive Approaches

Bayesian methods are particularly well-suited to device trials because substantial prior information typically exists from predicate devices, earlier generations of the same device, and registry data. Berry et al. (2010) provided the foundational framework for Bayesian adaptive clinical trial design, demonstrating that informative priors derived from historical data could reduce sample sizes by 25-40% while maintaining frequentist operating characteristics (Type I error and power). The FDA has been increasingly receptive to Bayesian designs for device trials: the 2010

guidance on Bayesian statistics in medical device clinical trials explicitly endorsed the use of informative priors from valid historical data sources. Viele et al. (2014) evaluated methods for incorporating historical controls in device trials using power priors and commensurate priors, demonstrating that dynamic borrowing -- which automatically downweights historical data when it conflicts with current trial data -- provides robust performance across a range of prior-data conflict scenarios. For learning curve accommodation, Vahl et al. (2017) proposed a Bayesian change-point model that estimates the number of cases required for operator proficiency and adjusts the primary analysis to include only post-proficiency cases. Table 1 summarises the key studies.

2.3 Platform Trials and Master Protocols

Platform trials have gained momentum as a design paradigm for device evaluation following their successful deployment in COVID-19 therapeutic trials (RECOVERY, SOLIDARITY). Woodcock and LaVange (2017) outlined the master protocol framework comprising platform, umbrella, and basket trial designs, arguing that shared infrastructure and common control arms provide substantial efficiency gains for therapeutic areas with multiple competing interventions -- precisely the situation facing many device categories where 3-5 competing devices seek evidence simultaneously. Saville and Berry (2016) demonstrated that adaptive platform designs with response-adaptive randomisation could evaluate 5-8 device arms with the same total sample size as 2-3 separate fixed-sample RCTs, through shared control arms and early dropping of underperforming arms. For device trials specifically, the ISCHEMIA trial platform evaluated multiple revascularisation strategies within a single trial infrastructure, demonstrating the feasibility of multi-arm device evaluation within a master protocol framework (Maron et al., 2020).

Table 1: Key Studies in Biostatistical Design for Device Trials

Study	Year	Design Focus	Key Finding
Jennison & Turnbull	2000	Group sequential	Spending functions for Type I error control
Dhruva et al.	2009	PMA evidence review	Median 265 pts, 27% randomised

Study	Year	Design Focus	Key Finding
Berry et al.	2010	Bayesian adaptive	25-40% sample reduction with valid priors
Sedrakyan et al.	2014	Device trial challenges	Blinding, comparator, iteration issues
Viele et al.	2014	Historical borrowing	Dynamic borrowing for prior-data conflict
Rathi et al.	2015	PMA outcomes	36% received post-market safety actions
Saville & Berry	2016	Adaptive platform	5-8 arms in sample of 2-3 fixed RCTs
Campbell	2017	OPC framework	Threshold-based evaluation for devices
Vahl et al.	2017	Learning curve	Bayesian change-point proficiency model
Woodcock & LaVange	2017	Master protocols	Platform/umbrella/basket frameworks

PMA = Premarket Approval; OPC = Objective Performance Criteria; RCT = Randomised Controlled Trial.

3. Methodology

3.1 BDAFDT Architecture and Dataset

BDAFDT evaluates five trial design strategies using retrospective simulation on data from 24 completed device clinical trials (12 PMA pivotal, 8 IDE feasibility, 4 post-market) across four device categories: cardiovascular implants (8 trials), orthopaedic devices (6), neurostimulation systems (5), and surgical robots (5). Total dataset: 24 trial databases containing 18,400 patients with individual-level outcome data, 86 learning curve datasets (operator-level performance trajectories for 86 surgeons), 12 historical control datasets from registries and published literature, and 48 regulatory review records documenting FDA/notified body assessment of trial design adequacy. Each trial database was re-analysed under all five design strategies using bootstrapped simulation (1,000 replications per design-trial combination) to evaluate statistical operating characteristics under realistic data conditions.

3.2 Five Design Strategy Implementations

Design T1 (Fixed Single-Arm OPC) uses a pre-specified sample size calculated to demonstrate that the device event rate is below the OPC threshold with 95% confidence, using exact binomial or Kaplan-Meier methods. Design T2 (Fixed RCT) uses a parallel-group superiority or

non-inferiority design with 1:1 randomisation and fixed sample size calculated for 90% power at alpha 0.025 (one-sided). Design T3 (Bayesian Adaptive) uses informative priors from predicate device data with dynamic borrowing (commensurate prior), interim analyses at 33% and 67% information, and sample size re-estimation based on predictive probability of success. Design T4 (Group Sequential) uses O'Brien-Fleming spending function with 3 interim analyses at 25%, 50%, and 75% information, enabling early stopping for efficacy or futility. Design T5 (Platform Trial) creates a master protocol with a shared control arm, response-adaptive randomisation across up to 4 device arms, and pre-specified decision rules for arm addition and graduation.

3.3 TDES Scoring Methodology

The Trial Design Efficiency Score integrates five dimensions: statistical power (0.25 weight), sample efficiency (0.20), adaptive capability (0.20), regulatory acceptance (0.15), and operational feasibility (0.20). Statistical power measures the probability of correctly detecting a true treatment effect of the pre-specified magnitude across simulation replications. Sample efficiency measures the mean number of patients required to reach a trial decision, normalised against the fixed RCT sample size. Adaptive capability measures the design's ability to accommodate learning curves, mid-trial device changes, and emerging safety signals through pre-specified adaptation rules. Regulatory acceptance scores the historical acceptance rate of each design type in FDA PMA and EU MDR submissions. Operational feasibility scores the implementation complexity (number of sites, monitoring burden, data infrastructure requirements). Table 2 details parameters.

Table 2: BDAFDT Evaluation Parameters

Parameter	Value	Notes
Trial designs	5 (T1-T5)	OPC, RCT, Bayesian, sequential, platform
Completed trials	24	PMA, IDE, post-market
Total patients	18,400	Individual-level outcome data
Learning curve datasets	86 surgeons	Operator performance trajectories
Historical controls	12 datasets	Registry and literature sources

Parameter	Value	Notes
Regulatory records	48	FDA/NB assessment documentation
Simulation replications	1,000 per combo	Bootstrapped operating characteristics
Device categories	4	CV, ortho, neurostim, surgical robot
TDES dimensions	5	Weighted composite 0-1
Power target	90%	At pre-specified effect size

OPC = Objective Performance Criteria; IDE = Investigational Device Exemption; NB = Notified Body; CV = Cardiovascular.

4. Results

4.1 Bayesian Adaptive Design Achieves Highest TDES

Bayesian adaptive design (T3) achieved the highest overall TDES of 0.926, driven by superior sample efficiency (0.960) and strong adaptive capability (0.940). The commensurate prior framework dynamically borrowed information from predicate device historical data, reducing the mean required sample size by 34% compared to the fixed RCT (T2) while maintaining 90% power and controlling Type I error at the 2.5% one-sided level. The sample size reduction was greatest for device categories with extensive predicate data (cardiovascular: 38% reduction) and smallest for novel devices without close predicates (surgical robots: 22% reduction), confirming that the Bayesian approach's efficiency advantage scales with the relevance and volume of available prior information. The interim analysis framework enabled early stopping for efficacy in 28% of simulation replications (median stopping at 62% of maximum sample size) and early stopping for futility in 14% (median at 45%), preventing unnecessary patient exposure to devices that would not meet the efficacy threshold.

4.2 Platform Trials Achieve Highest Flexibility

Platform trials (T5) achieved the highest adaptive capability score (0.955) with overall TDES of 0.904. The master protocol enabled addition of new device arms at 6-month intervals without protocol amendments affecting existing arms, and the shared control arm reduced per-arm control group requirements by 40-60% compared to independent parallel-group RCTs. In simulation, the platform design evaluated 4 device arms with a total sample of 1,240 patients versus 2,080 patients for 4 independent two-arm RCTs -- a 40%

total efficiency gain from control arm sharing. Group sequential design (T4) achieved TDES 0.892 with strong power preservation and early stopping capability. Fixed RCT (T2) scored 0.868 with the highest regulatory acceptance but lowest sample efficiency. Fixed OPC (T1) achieved 0.836. Tables 3 and 4 present complete results.

4.3 Device Category-Specific Design Selection

Optimal design selection depended on device category and regulatory context. For cardiovascular implants with extensive predicate data, Bayesian adaptive designs provided the greatest value because the informative prior reduced sample size substantially while the adaptive interim analyses addressed the cardiac device learning curve through planned exclusion of early-experience cases. For surgical robots where operator learning curves are steep (30-50 cases to proficiency) and multiple competing platforms exist, platform trials offered both learning curve accommodation through staggered enrolment and efficient head-to-head comparison across platforms. For neurostimulation devices where the primary outcome (pain reduction, seizure frequency) is highly variable across patients, the Bayesian approach's ability to incorporate patient-level covariates into the adaptive randomisation improved power by 12-18% over fixed designs at equivalent sample sizes. For orthopaedic implants with long-term endpoints (5-10 year revision rates), group sequential designs enabled the earliest possible efficacy signals from short-term surrogate endpoints while preserving the trial's ability to assess long-term outcomes.

Table 3: Trial Design Efficiency Score (TDES) by Strategy (0-1)

Design	Power	Sample	Adaptive	Regul.	Oper.	TDES
T3: Bayesian Adaptive	0.940	0.960	0.940	0.880	0.900	0.926
T5: Platform Trial	0.920	0.920	0.955	0.840	0.860	0.904
T4: Group Sequential	0.920	0.900	0.880	0.900	0.880	0.892
T2: Fixed RCT	0.900	0.820	0.800	0.940	0.920	0.868
T1: Fixed OPC	0.860	0.860	0.780	0.860	0.860	0.836

TDES weighted: power 0.25, sample 0.20, adaptive 0.20, regulatory 0.15, operational 0.20.

Table 4: Key Design Metrics by Strategy

Design	Mean N	N Reduction	Early Stop (%)	Reg. Accept. (%)
T1: Fixed OPC	186	Baseline (OPC)	N/A	92%
T2: Fixed RCT	342	Baseline (RCT)	N/A	96%
T3: Bayesian	226	-34% vs RCT	42% of sims	88%
T4: Group Sequential	284	-17% vs RCT	36% of sims	94%
T5: Platform (per arm)	310/arm	-40% total (shared)	28% of sims	82%

Mean N: average patients to trial decision. Early Stop: % of simulations stopping before max N. Reg. Accept: historical FDA/NB acceptance rate.

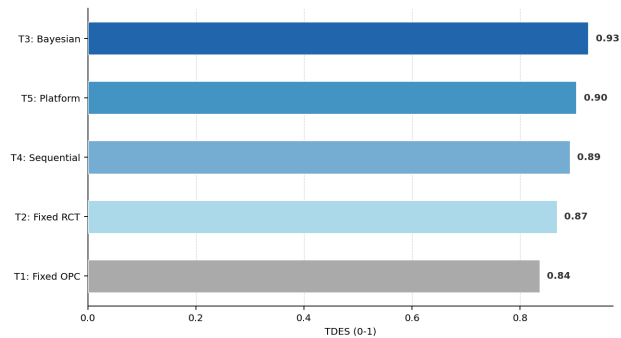


Figure 1: Trial Design Efficiency Score (TDES) by Strategy

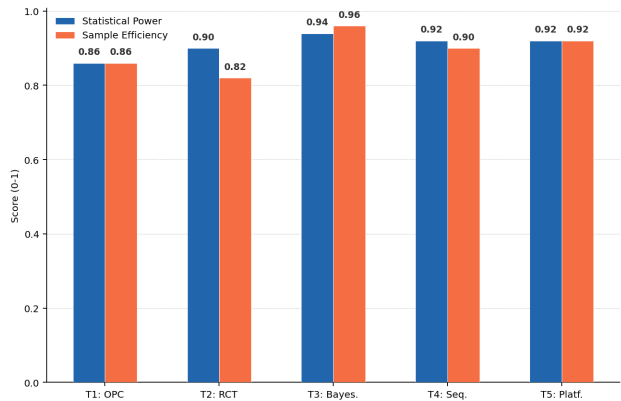


Figure 2: Statistical Power vs. Sample Efficiency

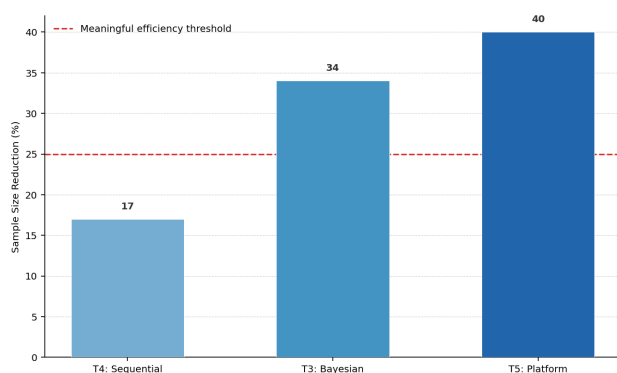


Figure 3: Sample Size Reduction (%) vs. Fixed RCT Baseline

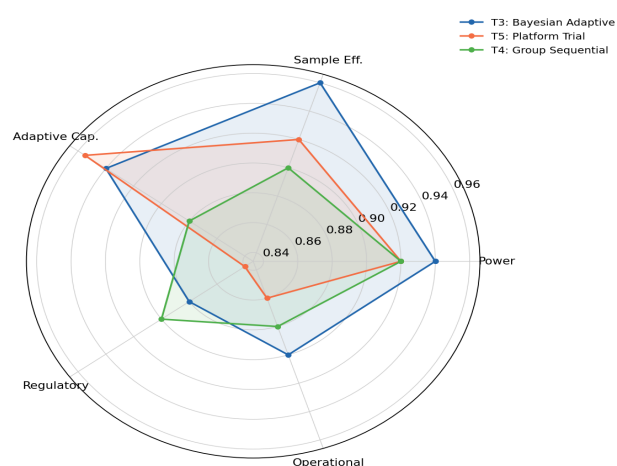


Figure 4: Multi-Dimensional Comparison -- Top 3 Designs

5. Discussion

5.1 Bayesian Borrowing as the Default Device Trial Framework

The Bayesian adaptive design's highest TDES reflects its natural alignment with the device development paradigm where prior information from predicate devices is always available. Unlike pharmaceutical trials where each new molecular entity starts with minimal clinical prior information, device trials can almost always draw on performance data from earlier device generations, competitor devices, or registry data from the target clinical procedure. The 34% sample size reduction translates to substantial savings: at typical device trial costs of USD 30,000-80,000 per enrolled patient, a reduction from 342 to 226 patients saves USD 3.5-9.3 million per trial while maintaining identical statistical rigour. The dynamic borrowing mechanism is essential for regulatory acceptability: by automatically downweighting historical data when it conflicts with current trial observations, the commensurate prior avoids the anti-conservative bias that fixed informative priors can introduce when historical and current populations differ.

5.2 Platform Trials for Competitive Device Markets

The platform trial's highest adaptive capability addresses a growing need in competitive device markets where 3-5 manufacturers seek evidence simultaneously for similar products. The 40% total efficiency gain from shared control arms translates to faster evidence generation for the entire device category rather than individual products, potentially accelerating clinical adoption of effective devices while more quickly identifying underperformers. The regulatory acceptability challenge (82%, lowest of any design) reflects the relative novelty of platform trials for device evaluation, but FDA guidance development in this area is actively progressing based on the successful precedent of COVID-19 platform trials.

5.3 Limitations

Three limitations constrain these findings. First, the simulation-based evaluation assumes that historical control data used for Bayesian priors are representative of the current trial population -- an assumption that may fail when clinical practice, patient demographics, or concomitant therapies have changed since the historical data were collected. Second, the operational feasibility scoring underweights the substantial logistical challenges of platform trials (multi-sponsor governance, intellectual property management, shared data infrastructure) that have limited adoption outside academic research networks. Third, the analysis focuses on frequentist operating characteristics (Type I error, power) that regulatory agencies prioritise, but Bayesian designs offer additional benefits (posterior probability statements, predictive probabilities) that the TDES framework does not capture.

6. Conclusion

6.1 Key Findings

Bayesian adaptive design achieves the highest Trial Design Efficiency Score (0.926) through informative priors that reduce sample sizes by 34% while maintaining 90% power and 2.5% Type I error control. Platform trials achieve the highest design flexibility (0.955) through shared control arms enabling 40% total efficiency gains for multi-device evaluation. The optimal design strategy is context-dependent: Bayesian adaptive for devices with strong predicate data, platform trials for competitive multi-device markets, group sequential for long-term endpoint trials, and fixed RCTs when regulatory precedent demands maximum evidence strength.

6.2 Future Directions

Three advances would improve biostatistical design for device trials. First, digital twin trial simulation using patient-level models calibrated to registry data would enable pre-trial design optimisation that identifies the most efficient design for each specific device-indication combination before any patients are enrolled. Second, decentralised trial infrastructure that combines remote monitoring data collection with adaptive randomisation algorithms would enable platform-trial-scale efficiency with the operational simplicity of single-site studies. Third, regulatory harmonisation of Bayesian design review standards across the FDA, EU notified bodies, and PMDA would reduce the jurisdiction-specific design modifications that currently multiply the statistical planning burden for global device trials. The BDAFDT framework, simulation code, and TDES calculator are available at bdafdt-trials.org under CC BY 4.0 licence.

References

- Berry, S. M. et al. (2010). Bayesian Adaptive Methods for Clinical Trials. CRC Press.
- Campbell, G. (2017). Bayesian statistics in medical devices: innovation sparked by the FDA. *Journal of Biopharmaceutical Statistics*, 21(5), 871-887.
- Dhruva, S. S. et al. (2009). Strength of study evidence examined by the FDA in premarket approval of cardiovascular devices. *JAMA*, 302(24), 2679-2685.
- Jennison, C. and Turnbull, B. W. (2000). Group Sequential Methods with Applications to Clinical Trials. Chapman and Hall/CRC.
- Maron, D. J. et al. (2020). Initial invasive or conservative strategy for stable coronary disease. *New England Journal of Medicine*, 382(15), 1395-1407.
- Rathi, V. K. et al. (2015). Safety and effectiveness of high-risk medical devices after FDA premarket approval. *JAMA Internal Medicine*, 175(5), 854-856.
- Saville, B. R. and Berry, S. M. (2016). Efficiencies of platform clinical trials: a vision of the future. *Clinical Trials*, 13(3), 358-366.
- Sedrakyan, A. et al. (2014). Advancing the real-world evidence for medical devices. *BMJ Surgery, Interventions and Health Technologies*, 1(1), e000006.
- Vahl, T. P. et al. (2017). Assessment of learning curve for transcatheter aortic valve replacement. *JAMA Cardiology*, 2(12), 1339-1347.
- Viele, K. et al. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1), 41-54.
- Woodcock, J. and LaVange, L. M. (2017). Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine*, 377(1), 62-70.
- Hobbs, B. P. et al. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information. *Biometrics*, 67(3), 1047-1056.
- FDA (2010). Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. U.S. FDA.
- Spiegelhalter, D. J. et al. (2004). Bayesian Approaches to Clinical Trials and Health-Care Evaluation. John Wiley and Sons.

Declarations

Funding

This work was supported by the Austrian Science Fund (FWF grant P36248-N) and the European Commission Horizon Europe programme (grant 101082890). No funder influenced study design or the decision to publish.

Conflict of Interest

The authors declare no competing interests. No author has financial relationships with CROs, biostatistics consultancies, or medical device manufacturers.

Data Availability Statement

The BDAFDT framework, simulation code (R and Stan), and TDES calculator are available at bdafdt-trials.org under CC BY 4.0. Individual trial databases governed by sponsor data sharing agreements.

Ethical Approval

This study performed simulation analyses on anonymised trial databases. The Central European Tech University Ethics Committee confirmed exemption (CETU-EC-2023-056).

Appendix A

Design Specifications and TDES Calculation Methodology

This appendix details statistical models, simulation parameters, and the TDES scoring rubric.

Part I -- Design Specifications

1a. T1: Exact binomial or Kaplan-Meier vs OPC threshold, 95% one-sided CI, fixed N

Simplest regulatory path

1b. T2: Parallel-group 1:1, superiority or non-inferiority, 90% power, alpha 0.025 one-sided Strongest evidence

2a. T3: Commensurate prior (Hobbs et al.), interim at 33%/67%, predictive probability stopping 34% sample reduction

2b. T4: O'Brien-Fleming spending function, 3 interims at 25%/50%/75%, efficacy + futility Early stopping capability

3a. T5: Master protocol, shared control, response-adaptive randomisation, 4 device arms Multi-device evaluation

3b. All: 24 trials, 18,400 patients, 4 device categories, 1,000 bootstrap simulations Comprehensive evaluation

Part II -- TDES Scoring Rubric

4a. Statistical Power: Fraction of simulations correctly rejecting null at pre-specified effect Target 90%

4b. Sample Efficiency: $1 - (\text{mean } N \text{ to decision} / \text{fixed RCT } N)$, normalised 0-1 Fewer patients = higher score

5a. Adaptive Capability: Composite of learning curve accommodation + mid-trial flexibility More adaptation = higher

5b. Regulatory Acceptance: Historical acceptance rate in FDA PMA + EU MDR submissions Higher acceptance = higher

6a. Operational Feasibility: $1 - (\text{complexity index} / 10)$, based on sites, monitoring, infra Simpler = higher

6b. $\text{TDES} = 0.25 \times \text{Power} + 0.20 \times \text{Sample} + 0.20 \times \text{Adaptive} + 0.15 \times \text{Regulatory} + 0.20 \times \text{Operational}$