zetyra | Evidence in the Wild

Zetyra: A Validated Suite of Statistical Calculators for Efficient Clinical Trial Design

Technical White Paper

Version 1.0 | January 2026

Lu (Maggie) Qian, MS
Evidence in the Wild, Zetyra

maggie@zetyra.com
https://zetyra.com
Validation Repository: https://github.com/evidenceinthewild/zetyra-validation

Contents

1 Executive Summary

FDA Published Bayesian Guidance
On January 12, 2026, FDA released draft guidance extending Bayesian methodology to drugs and biologics (satisfying PDUFA VII commitment). **FDA Commissioner Marty Makary:** "Bayesian methodologies help address two of the biggest problems of drug development: high costs and long timelines." This white paper demonstrates exactly this value proposition through validated calculators and quantified case studies.

The Challenge: Phase III clinical trials in oncology and cardiology average \$50–100 million[1] and require 4–6 years from first patient to database lock[2] (DiMasi et al., 2016; Moore et al., 2018). Conservative statistical designs—failing to leverage baseline covariates, fixed-sample approaches without interim monitoring, and frequentist paradigms for Phase II decisions—often inflate sample sizes by 15–35%[3] relative to efficient alternatives. Three proven methodologies can substantially reduce trial costs and duration, but existing software packages are expensive (\$5,000–\$15,000 annually[4]), complex to deploy, and lack transparent validation.

The Solution: Zetyra is a web-based platform offering three validated statistical calculators that enable efficient clinical trial design. Key differentiators:

- **Comprehensive methodology:** Integrates CUPED, group sequential, and Bayesian methods in single platform
- **Affordable:** Monthly subscription vs. \$5K–\$15K perpetual licenses
- **Transparent validation:** Public validation suite (51 automated tests) vs. proprietary validation approaches
- **Accessible:** Web-based interface vs. IT department installation requirements
- **Accurate:** Maximum deviation 0.0046 z-score[5] (O'Brien-Fleming boundaries, K=2–5 looks, one-sided $\alpha = 0.025$) vs. our pre-specified acceptance criterion of ± 0.05 z-score

1.1 Key Results from Validation

- **Group Sequential Design:** 30 tests passed, maximum deviation 0.0046 z-score when benchmarked against gsDesign R package (gold standard)
- **CUPED:** 12 tests passed, exact match to analytical variance reduction formula (VRF = $1 - \rho^2$)

[1] Moore et al. (2018) analyzed 138 pivotal trials supporting FDA approvals in 2015-2016, finding median Phase III cost of \$19 million (range \$12M–\$33M for trials with 50–300 patients). Oncology and cardiovascular trials with 500–2,000+ patients commonly exceed \$50–100M based on per-patient costs of \$40K–\$60K (Sertkaya et al., 2016).

[2] DiMasi et al. (2016) report median clinical development time of 6–7 years from IND to NDA submission. Phase III trials typically require 2–3 years enrollment plus 1–3 years follow-up depending on endpoint.

[3] For continuous outcomes with baseline covariate correlation ρ, failure to adjust inflates sample size by factor $1/(1-\rho^2)$. With $\rho = 0.4$–0.6 (typical per Walters et al., 2019), this yields 1.19–1.56 inflation factor (19–56% increase), or equivalently 16–36% reduction potential.

[4] Based on published list prices as of January 2026: East (\$15,000/year), PASS (\$7,995/year), ADDPLAN (\$10,000–\$12,000/year), nQuery (\$5,995/year).

[5] The 0.0046 z-score represents the maximum absolute difference $|Z_{\text{Zetyra}} - Z_{\text{gsDesign}}|$ across 23 boundary comparisons. Coverage: O'Brien-Fleming and Pocock spending functions, K=2,3,4,5 interim analyses, one-sided $\alpha = 0.025$, 90% power.

- **Bayesian:** 9 tests passed, exact match to conjugate prior analytical solutions

1.2 Business Impact (Representative Examples)

A Phase II oncology trial (240 patients standard design) can potentially be reduced to:

- **168 patients with CUPED** (30% reduction based on $\rho = 0.55$ correlation, assuming \$50K per-patient cost yields \$3.6M savings; see Case Study 7.1 for detailed assumptions)
- **Expected 20–40% sample size reduction with group sequential design** under alternative hypothesis (varies by boundary type and number of looks; Jennison & Turnbull, 2000)
- **Improved go/no-go decision-making with Bayesian monitoring** (quantitative risk assessment vs. binary p-value thresholds; see Case Study 7.3)

1.3 Target Markets

- **Primary:** Biotech and pharmaceutical companies (Series B+) designing Phase II/III trials
- **Secondary:** Academic medical centers conducting investigator-initiated trials
- **Tertiary:** CROs providing statistical design services

1.4 Reading Guide by Audience

- **Executives & Decision-Makers:** Sections 1 (Executive Summary), 2 (Introduction), 7 (Case Studies), 8 (Conclusion)
- **Biostatisticians & Technical Reviewers:** Sections 3–6 (CUPED, GSD, Bayesian methodology, Validation)
- **Regulatory & Quality Professionals:** Sections 1.4 (Intended Use), 5.5 (Regulatory Context), 6 (Validation Framework)

This white paper provides comprehensive technical documentation of Zetyra's methodology, validation framework, and real-world applications for senior biostatisticians evaluating clinical trial design software.

1.5 Intended Use and Scope

Intended Use Statement

Zetyra calculators are intended for:

- Sample size and power calculations for clinical trial planning
- Generating statistical parameters (boundaries, information fractions, predictive probabilities) as inputs to statistical analysis plans
- Supporting regulatory submission documentation with validated calculations
- Educational and training purposes in clinical trial design

Zetyra calculators are NOT intended for:

- Making final regulatory or clinical decisions without biostatistician review
- Replacing sponsor quality systems or validation requirements
- Direct execution of interim analyses (calculators provide boundaries; execution requires separate clinical trial management systems)
- Medical device or diagnostic classification decisions

User responsibility: Zetyra provides validated calculation tools. Sponsors remain responsible for: (1) verifying fit-for-purpose within their quality management system, (2) appropriate application of methodologies to specific trial contexts, and (3) regulatory strategy decisions.

2 Introduction

2.1 The Clinical Trial Efficiency Problem

Clinical development of new therapeutics represents one of the most capital-intensive endeavors in modern medicine. DiMasi et al. (2016) estimated the capitalized cost to bring a single drug from discovery through FDA approval at $2.6 billion, with Phase II and Phase III trials accounting for approximately 60% of total development costs. Moore et al. (2018) analyzed 138 pivotal trials supporting novel therapeutic agents approved by FDA in 2015–2016, finding median Phase III costs of $19 million (range: $12M–$33M for trials with 50–300 patients). Oncology and cardiovascular trials at the upper end of enrollment (500–2,000+ patients) commonly exceed $50–100 million (Sertkaya et al., 2016).

Conservative design practices systematically inflate these already-substantial costs. Three common inefficiencies dominate:

1. **Failure to leverage baseline covariates:** Standard power calculations ignore correlations ($\rho = 0.4$–0.7 typical for many endpoints; Walters et al., 2019) between baseline measurements and treatment outcomes. For continuous outcomes, failing to adjust for baseline covariates inflates sample sizes by a factor of $1/(1-\rho^2)$, yielding 15–35% overestimation when ρ ranges from 0.4 to 0.6 (Frison & Pocock, 1992; Teerenstra et al., 2012).
2. **Fixed-sample designs despite interim data:** Most trials continue to planned comple-

tion despite accumulating interim evidence of efficacy or futility. Group sequential designs with pre-specified stopping boundaries can reduce expected sample size under the alternative hypothesis by 15–30% (O'Brien-Fleming) to 30–40% (Pocock), with commensurate reductions in expected trial duration (Jennison & Turnbull, 2000).

3. **Frequentist paradigm for Phase II go/no-go decisions:** Traditional hypothesis tests provide binary answers ($p < 0.05$ or not) without quantifying the probability of Phase III success given Phase II data. Bayesian predictive probability frameworks enable more nuanced decisions; simulation studies suggest well-calibrated priors can reduce false-go rates relative to p-value thresholds, though the magnitude depends heavily on prior specification and decision thresholds (Berry et al., 2010; Spiegelhalter & Freedman, 1986).

Existing software solutions are inadequate:

Limitation	Impact on Adoption
High cost: \$5K–\$15K/year	Small biotechs (Series A/B) priced out
IT barriers: Desktop install, version control	Requires IT department involvement
Limited scope: Separate tools per method	Users must purchase multiple products
Opaque validation: No published benchmarks	Public, independently reproducible test suites not typically provided
Poor documentation: Sparse regulatory citations	Additional work for FDA/EMA submissions

Table 1: Limitations of existing clinical trial design software

2.2 Zetyra Platform Overview

Zetyra addresses these inefficiencies through three integrated, validated statistical calculators accessible via a web-based platform:

1. CUPED (Controlled-Experiment Using Pre-Experiment Data):

- Calculates sample size reduction from baseline covariate adjustment
- Inputs: Standard design parameters + baseline-outcome correlation (ρ)
- Outputs: Variance reduction factor ($1 - \rho^2$), adjusted sample size, expected power gain
- Use cases: Any continuous or time-to-event endpoint with measured baseline

2. Group Sequential Design:

- Calculates stopping boundaries for interim analyses
- Supports: O'Brien-Fleming, Pocock, and alpha-spending function boundaries
- Outputs: Z-score boundaries, sample sizes at each look, expected sample size under H_0 and H_1
- Use cases: Phase II/III trials with interim DSMB reviews

3. Bayesian Predictive Power:

- Calculates probability of trial success given interim data

- Supports: Beta-binomial (binary endpoints), normal-normal (continuous endpoints)
- Outputs: Posterior distribution, predictive probability, futility/graduation thresholds
- Use cases: Phase II go/no-go decisions, adaptive dose-finding

Platform Features:

- **Web-based:** Zero installation, instant updates, works on any device
- **Integrated:** Shared parameter sets, consistent UI across calculators
- **Transparent:** Public validation suite with 51 automated tests
- **Regulatory-ready:** FDA/EMA guidance citations embedded in outputs
- **API-enabled:** RESTful endpoints for programmatic access (Appendix A)
- **Monthly subscription:** \$99–\$299/month vs. \$5K–\$15K/year perpetual licenses

2.3 Document Purpose and Scope

This white paper provides comprehensive technical documentation for biostatisticians evaluating Zetyra for clinical trial design. Sections 3–5 present detailed methodology for each calculator, including:

- Theoretical foundations and mathematical derivations
- Implementation algorithms and numerical considerations
- Regulatory guidance and FDA/EMA citations
- Practical application scenarios and decision frameworks

Section 6 describes the validation framework, presenting results from 51 automated tests benchmarking Zetyra against:

- **gsDesign R package** (gold standard for group sequential design)
- **Analytical formulas** (closed-form solutions for CUPED and Bayesian methods)
- **Published clinical trials** (HPTN 083, HeartMate II design replications)

Section 7 provides three detailed case studies demonstrating real-world applications and quantifying cost/time savings. Section 8 synthesizes conclusions and describes the product roadmap.

3 CUPED: Covariate-Adjusted Power Analysis

3.1 Theoretical Foundation

CUPED (Controlled-experiment Using Pre-Experiment Data) is a variance reduction technique that leverages baseline covariates to improve statistical power in randomized controlled trials. Originally developed by Microsoft Research (Deng et al., 2013) for online A/B testing, CUPED has proven applications in clinical trial design where baseline measurements correlate with treatment outcomes.

3.1.1 Relationship to ANCOVA

CUPED is fundamentally related to Analysis of Covariance (ANCOVA), a classical statistical method dating to Fisher (1932). While ANCOVA adjusts for baseline covariates in the analysis stage, CUPED extends this principle to the design stage, enabling more efficient sample size planning. The key insight is that if a baseline measurement X is correlated with the outcome Y, incorporating X into the analysis reduces unexplained variance and increases statistical power.

Mathematically, consider a two-arm randomized trial comparing treatment vs. control on continuous outcome Y. Standard analysis compares means: $\bar{Y}_{\text{treatment}} - \bar{Y}_{\text{control}}$. ANCOVA/CUPED instead compares adjusted means that account for baseline covariate X:

Adjusted estimator:

$$\boxed{Y_{\text{CUPED}} = Y - \theta(X - \mathbb{E}[X])} \tag{1}$$

where θ is the optimal adjustment coefficient (derived below). This adjustment removes the component of Y that is predictable from X, leaving only the residual variance. Because randomization ensures X is balanced across treatment arms, $\mathbb{E}[X]$ is identical in both groups, preserving unbiased treatment effect estimation while reducing variance.

3.1.2 Connection to Control Variates

CUPED derives from control variates in Monte Carlo simulation (Lavenberg & Welch, 1981). The key insight: if a baseline covariate X correlates with outcome Y, we can construct a variance-reduced estimator by subtracting the predictable component.

The adjusted estimator $Y_{\text{CUPED}} = Y - \theta(X - \mathbb{E}[X])$ is unbiased regardless of the relationship between X and Y (because $\mathbb{E}[X - \mathbb{E}[X]] = 0$), but variance reduction is maximized when the relationship is linear with coefficient $\theta = \text{Cov}(X, Y)/\text{Var}(X)$. This is equivalent to the OLS regression coefficient.

3.1.3 Historical Context

Variance reduction through baseline covariate adjustment has been studied extensively in biostatistics:

- **Frison & Pocock (1992):** Demonstrated 15–35% sample size reductions for repeated measures trials with $\rho = 0.4$–0.6
- **Vickers & Altman (2001):** Advocated baseline adjustment for continuous outcomes in RCTs
- **Teerenstra et al. (2012):** Extended methods to cluster randomized trials
- **Deng et al. (2013):** Formalized CUPED for online experimentation with pre-experiment data

3.2 Mathematical Framework

3.2.1 Optimal Adjustment Coefficient

The adjustment coefficient θ that minimizes $\text{Var}(Y_{\text{CUPED}})$ is derived via straightforward calculus:

$$\boxed{\theta^* = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}} \tag{2}$$

This is mathematically equivalent to the ordinary least squares (OLS) regression coefficient from regressing Y on X. Substituting θ^* into the variance expression:

$$\text{Var}(Y_{\text{CUPED}}) = \text{Var}(Y) - \frac{[\text{Cov}(X, Y)]^2}{\text{Var}(X)} = \text{Var}(Y) \times (1 - \rho^2) \tag{3}$$

where ρ is the Pearson correlation coefficient between X and Y:

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \times \text{Var}(Y)}} \tag{4}$$

3.2.2 Variance Reduction Factor (VRF)

The variance reduction factor quantifies the proportional decrease in outcome variance:

$$\boxed{\text{VRF} = \frac{\text{Var}(Y_{\text{CUPED}})}{\text{Var}(Y)} = 1 - \rho^2} \tag{5}$$

Key Property: VRF depends only on the squared correlation ρ^2. This has important implications:

Correlation (ρ)	VRF	Variance Reduction	Sample Size Reduction
0.0 (no correlation)	1.00	0%	0%
0.5 (moderate)	0.75	25%	25%
0.7 (strong)	0.51	49%	49%
0.9 (very strong)	0.19	81%	81%

Table 2: Variance reduction as a function of baseline-outcome correlation

3.2.3 Sample Size Adjustment Formula

For a two-sample t-test comparing treatment arms with equal allocation, standard sample size is:

$$n_{\text{standard}} = \frac{2\sigma^2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\delta^2} \tag{6}$$

where:

- σ^2 = outcome variance
- δ = treatment effect (difference in means)
- $Z_{1-\alpha/2}$ = critical value for two-sided test (1.96 for $\alpha = 0.05$)
- $Z_{1-\beta}$ = critical value for power (0.84 for 80% power, 1.28 for 90% power)

With CUPED adjustment:

$$\boxed{n_{\text{CUPED}} = n_{\text{standard}} \times (1 - \rho^2) = n_{\text{standard}} \times \text{VRF}} \tag{7}$$

Interpretation: Sample size decreases proportionally to variance reduction. A correlation of $\rho = 0.6$ (VRF = 0.64) reduces required sample size by 36%.

3.2.4 Design Factor

Some authors (Frison & Pocock, 1992) express efficiency gain as a "design factor":

$$\text{Design Factor} = \frac{n_{\text{CUPED}}}{n_{\text{standard}}} = 1 - \rho^2 \tag{8}$$

This directly translates correlation magnitude into sample size savings, facilitating communication with non-statisticians and budget planners.

3.2.5 Extension to Multiple Covariates

When multiple baseline covariates are available $(X_1, X_2, \ldots, X_p)$, CUPED generalizes to:

$$Y_{\text{CUPED}} = Y - \sum_{j=1}^{p} \theta_j (X_j - \mathbb{E}[X_j]) \tag{9}$$

The adjustment coefficients $(\theta_1, \ldots, \theta_p)$ are obtained from multiple regression of Y on $X_1, \ldots, X_p$. The variance reduction factor becomes:

$$\text{VRF} = 1 - R^2 \tag{10}$$

where R^2 is the multiple correlation coefficient (proportion of Y variance explained by all covariates). **Important:** This R^2 is computed from regressing outcome Y on baseline covariates $X_1, \ldots, X_p$ using pooled or blinded historical data, not including treatment indicators.

This can substantially exceed single-covariate VRF; for example, baseline tumor burden + performance status might achieve $R^2 = 0.50$–0.60 in oncology trials, compared to $\rho^2 = 0.30$–0.40 for tumor burden alone.

3.3 Implementation Details

3.3.1 Zetyra Implementation

Zetyra's CUPED calculator uses custom Python implementations built on NumPy and SciPy for numerical stability and computational efficiency. Key implementation features:

1. **Covariance Estimation:** Uses sample covariance when historical data are provided:

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{11}$$

2. **Variance Reduction Calculation:** Applies exact formula VRF $= 1 - \rho^2$ without approximation

3. **Sample Size Adjustment:** Multiplies standard power calculation by VRF, rounding up to ensure adequate power

4. **Sensitivity Analysis:** Offers correlation range inputs (e.g., $\rho = 0.4$–0.6) to assess robustness to estimation uncertainty

3.3.2 Estimating Baseline-Outcome Correlation

Practitioners have several options for estimating ρ:

Source	Advantages	Limitations
Historical trial data	Direct measurement	May not generalize
Published literature	Walters et al.: $\rho = 0.50$ median	Varies by endpoint
Pilot study	Population-specific	Small sample ($n = 30$–50)
External databases	Large sample sizes	Selection bias

Table 3: Sources for estimating baseline-outcome correlation

Walters et al. (2019) Benchmark Correlations:

Analysis of 464 correlations from 20 UK Health Technology Assessment trials:

- **Mean** $\rho = 0.50$ (median 0.51, SD 0.15, range -0.13 to 0.91)
- By outcome type:
  - Depression scales (PHQ-9): $\rho = 0.66$
  - Physical functioning (SF-36): $\rho = 0.64$
  - Quality of life (EQ-5D): $\rho = 0.55$
  - Pain scales (VAS): $\rho = 0.41$

Rule of thumb: For stable trait measures (depression, quality of life), assume $\rho = 0.60$–0.70. For state/symptomatic measures (pain, fatigue), assume $\rho = 0.40$–0.50.

3.4 Regulatory Considerations

3.4.1 FDA Guidance for Industry: Adjusting for Covariates in Randomized Clinical Trials (May 2023)

The FDA released updated guidance explicitly encouraging covariate adjustment as "low-hanging fruit" to improve trial efficiency. Key provisions:

1. **Endorsement of Covariate Adjustment:**

   > "FDA encourages sponsors to consider covariate adjustment as a way to improve the precision of treatment effect estimates and increase statistical power."

2. **Applicable Settings:**

- Continuous outcomes: ANCOVA is recommended baseline-adjusted analysis
- Binary outcomes: Logistic regression adjusting for baseline covariates
- Time-to-event outcomes: Cox regression with baseline stratification

3. **Pre-Specification Requirement:**

   "Covariate adjustment should be pre-specified in the statistical analysis plan before database lock and unblinding."

4. **Covariate Selection:**

   - Baseline covariates only (not post-randomization variables)
   - Prognostic covariates correlated with outcome
   - Stratification factors (mandatory to adjust for)
   - Strong prognostic factors identified from literature

3.4.2 EMA Guideline on Adjustment for Baseline Covariates (February 2015)

The European Medicines Agency issued parallel guidance:

1. **Efficiency Gains:**

   "Adjustment for baseline covariates can lead to more precise estimation of the treatment effect and increased power, which is beneficial for trial efficiency."

2. **Balance vs. Imbalance:**

   - Adjustment beneficial even when randomization achieves good baseline balance
   - Greater benefit when chance imbalances occur (small trials)

3.4.3 ICH E9(R1): Addendum on Estimands and Sensitivity Analysis (November 2019)

ICH E9(R1) introduced the estimands framework, which clarifies the role of covariate adjustment:

1. **ANCOVA in Estimands Framework:**

   - ANCOVA provides conditional treatment effect (given covariate values)
   - Marginal treatment effect (population average) obtained by averaging over covariate distribution
   - Both are valid estimands; choice depends on inferential target

2. **Sensitivity Analyses:**

   - Unadjusted analysis as sensitivity check (should be consistent with adjusted)
   - Alternative covariate sets to assess robustness

3.5 Practical Applications

3.5.1 When to Use CUPED vs. Standard Power Calculations

Scenario	Recommendation	Rationale
Baseline measurement available	Use CUPED	Variance reduction increases power
Expected $\rho \geq 0.4$	Strong case for CUPED	$\geq 16\%$ sample size reduction
Small trial ($N < 100$)	Use CUPED	Greater impact of efficiency gains
Expensive endpoints	Use CUPED	Cost savings justify covariate collection
No baseline data	Standard calculation	Cannot adjust without covariate
Expected $\rho < 0.3$	Marginal benefit	$< 9\%$ sample size reduction

Table 4: Decision framework for CUPED vs. standard power calculations

3.5.2 When CUPED Doesn't Help (Limitations)

CUPED is not universally applicable. Consider these limitations:

- **Low correlation ($\rho < 0.3$):** Sample size reduction $< 9\%$; administrative overhead may exceed benefit
- **Unstable baseline measures:** If baseline measurement has high test-retest variability, observed ρ may overestimate true predictive value
- **Substantial missingness:** Missing baseline data reduces effective sample; complete-case analysis may introduce bias
- **Post-baseline covariates:** CUPED theory assumes pre-randomization covariates; post-baseline measures can introduce bias
- **Time-to-event endpoints:** Covariate adjustment for survival endpoints uses different methods (stratified Cox, adjusted log-rank); VRF formula applies to continuous outcomes
- **Binary endpoints:** Efficiency gains from covariate adjustment in logistic regression are generally smaller than for continuous outcomes

3.5.3 Estimating Correlation (ρ) for Planning

Reliable ρ estimates are critical for CUPED power calculations. Recommended approaches:

Source	Approach	Considerations
Historical trial data	Extract from prior studies	Best source if similar population/endpoint
Published literature	Meta-analysis of correlations	Walters et al. (2019) provides endpoint-specific data
Natural history studies	Observational cohort data	May overestimate if treatment affects trajectory
Blinded internal pilot	Estimate from first 20–30% of data	Most reliable; requires SAP pre-specification

Table 5: Approaches for estimating baseline-outcome correlation

Conservative defaults: When uncertain, use conservative ρ estimates. For stable trait measures (depression scales, quality of life), assume $\rho = 0.50$–0.60. For state/symptomatic measures (pain, fatigue), assume $\rho = 0.35$–0.45. Document assumptions in the statistical analysis plan.

3.5.4 Example Calculation with Interpretation

Scenario: Two-arm Phase III cardiovascular trial comparing Novel Drug vs. Placebo on change in 6-minute walk distance (6MWD).

Design Parameters:

- Effect size: $\delta = 30$ meters (clinically meaningful difference)
- Standard deviation: $\sigma = 80$ meters (from prior studies)
- Baseline-outcome correlation: $\rho = 0.55$ (baseline 6MWD predicts follow-up 6MWD)
- Significance: $\alpha = 0.05$ (two-sided)
- Power: 90% ($\beta = 0.10$)

Step 1: Standard Sample Size Calculation

$$n_{\text{standard}} = \frac{2 \times 80^2 \times (1.96 + 1.28)^2}{30^2} \tag{12}$$

$$= \frac{2 \times 6400 \times 10.50}{900} = \frac{134,400}{900} = 149.3 \approx 150 \text{ per arm} \tag{13}$$

Total $N = 300$ patients

Step 2: CUPED-Adjusted Sample Size

$$\text{VRF} = 1 - 0.55^2 = 1 - 0.3025 = 0.6975 \tag{14}$$

$$n_{\text{CUPED}} = 150 \times 0.6975 = 104.6 \approx 105 \text{ per arm} \tag{15}$$

Total $N_{\text{CUPED}} = 210$ patients

Step 3: Interpret Savings

- **Sample size reduction:** $300 - 210 =$ **90 patients (30%)**
- **Cost savings:** $90 \times \$40,000/\text{patient} =$ **\$3.6 million**
- **Timeline acceleration:** $90 \div 25$ patients/month = **3.6 months faster completion**

Conclusion: By leveraging baseline 6MWD measurement ($\rho = 0.55$), this trial reduces from 300 to 210 patients, saving \$3.6M and accelerating regulatory submission by ∼4 months. This efficiency gain requires proper pre-specification in the statistical analysis plan. ANCOVA adjustment is supported by FDA guidance (May 2023) and routinely used in cardiovascular trials.

4 Group Sequential Design

4.1 Theoretical Foundation

Group Sequential Designs (GSD) allow pre-planned interim analyses during clinical trials while maintaining overall Type I error control. This adaptive approach enables early termination for efficacy (if treatment effect is compelling) or futility (if success appears unlikely), substantially reducing expected trial duration and sample size compared to fixed-sample designs.

4.1.1 Historical Development

The foundations of sequential analysis date to World War II, when Abraham Wald (1947) developed sequential probability ratio tests for industrial quality control. Clinical trial applications emerged in the 1970s:

- **Pocock (1977):** Proposed constant boundaries across interim analyses
- **O'Brien & Fleming (1979):** Introduced conservative early boundaries that preserve final analysis significance level
- **Lan & DeMets (1983):** Generalized to alpha-spending functions, accommodating unequal interim spacing

4.1.2 Brownian Motion Theory

Group sequential theory rests on a fundamental result: under repeated sampling, the cumulative test statistic follows Brownian motion with drift. Specifically, let Z_k be the standardized test statistic at interim analysis k ($k = 1, \ldots, K$):

$$Z_k = \frac{\hat{\theta}_k}{\text{SE}(\hat{\theta}_k)} \tag{16}$$

where $\hat{\theta}_k$ is the treatment effect estimate and SE is standard error at analysis k. Under the null hypothesis (no treatment effect), $\{Z_k\}$ follows a Brownian motion with independent increments and variance proportional to information accrual.

4.1.3 Information Time

Sequential boundaries are defined in terms of **information time** (τ), the proportion of maximum statistical information available:

$$\tau_k = \frac{I_k}{I_K} \tag{17}$$

where I_k is information at analysis k and I_K is maximum (final analysis) information. For continuous outcomes with equal variance, information is proportional to sample size: $\tau_k = n_k/N$. For time-to-event outcomes, information equals number of events: $\tau_k = d_k/D$.

4.1.4 Type I Error Control

The fundamental challenge in interim monitoring is controlling familywise Type I error rate. Without adjustment, repeated testing inflates false positive probability:

Number of Looks (K)	Naive $\alpha = 0.05$ per look	True Type I Error
1	0.05	0.050
2	0.05	0.083
3	0.05	0.106
5	0.05	0.141

Table 6: Inflation of Type I error with multiple looks

To maintain overall $\alpha = 0.05$, boundaries must be adjusted. Group sequential methods achieve this through carefully calibrated critical values at each analysis.

4.2 Alpha-Spending Functions

4.2.1 Lan-DeMets Alpha-Spending Framework (1983)

Lan and DeMets introduced a flexible framework where the Type I error probability "spent" at each analysis is determined by an **alpha-spending function** $\alpha(\tau)$, which must satisfy:

1. $\alpha(0) = 0$ (no error spent before trial starts)
2. $\alpha(1) = \alpha$ (full error budget spent at final analysis)
3. $\alpha(\tau)$ is non-decreasing in τ

At analysis k with information time τ_k, the incremental alpha spent is:

$$\Delta\alpha_k = \alpha(\tau_k) - \alpha(\tau_{k-1}) \tag{18}$$

The critical value (boundary) Z_k is chosen such that the probability of crossing equals $\Delta\alpha_k$ under H_0.

4.2.2 O'Brien-Fleming Spending Function

The O'Brien-Fleming approach spends alpha conservatively early and more liberally later:

$$\boxed{\alpha_{\text{OF}}(t) = 2\left[1 - \Phi\left(\frac{Z_{\alpha/2}}{\sqrt{t}}\right)\right]} \tag{19}$$

where Φ is the standard normal CDF and t is information time.

Example for $K = 5$ looks, $\alpha = 0.025$ (one-sided):

Look	Info Time (t)	$\alpha(t)$ Cumulative	$\Delta\alpha$ Increment	Z-boundary
1	0.20	0.00001	0.00001	4.56
2	0.40	0.00055	0.00054	3.23
3	0.60	0.00385	0.00330	2.63
4	0.80	0.01096	0.00711	2.28
5	1.00	0.02500	0.01404	2.04

Table 7: O'Brien-Fleming boundaries for 5-look design

Interpretation: Very high thresholds early ($Z = 4.56$ at 20% information) require overwhelming evidence for early stopping. By final analysis, boundary approaches fixed-sample critical value (1.96).

4.2.3 Pocock Spending Function

Pocock boundaries spend alpha approximately equally at each look:

$$\alpha_{\text{Pocock}}(t) \approx \alpha \times t^{0.5} \tag{20}$$

This yields constant critical values across analyses (e.g., $Z \approx 2.41$ for $K = 5$, $\alpha = 0.025$).

4.2.4 Hwang-Shih-DeCani (HSD) Spending Function

A flexible parametric family that generalizes O'Brien-Fleming and Pocock:

$$\alpha_{\text{HSD}}(t; \gamma) = \alpha \times \frac{1 - e^{-\gamma t}}{1 - e^{-\gamma}} \tag{21}$$

- $\gamma = -4$: Approximates O'Brien-Fleming (conservative early)
- $\gamma = 0$: Linear spending
- $\gamma = 1$: Approximates Pocock (aggressive early)

4.2.5 Kim-DeMets Spending Functions

Another flexible family:

$$\alpha_{\text{KD}}(t; \rho) = \alpha \times t^{\rho} \tag{22}$$

- $\rho = 0$: O'Brien-Fleming type
- $\rho = 0.5$: Square root
- $\rho = 1$: Linear (Pocock-like)

4.2.6 Choosing a Spending Function

Spending Function	Alpha Allocation	When to Use
O'Brien-Fleming	Conservatively early, aggressively late	Preserve final p-value; pivotal trials
Pocock	Roughly equal across looks	Maximize early stopping probability
HSD ($\gamma < 0$)	Flexible conservative	Fine-tune between OF and Pocock
Linear	Proportional to information	Balanced approach

Table 8: Comparison of alpha-spending functions

FDA Preference: O'Brien-Fleming boundaries are most commonly used in regulatory submissions because they preserve statistical rigor—early stopping requires very strong evidence, and the final p-value is close to 0.05 if trial continues to completion.

4.3 Boundary Calculations

4.3.1 Efficacy Boundaries

Efficacy boundaries define Z-score thresholds for concluding treatment superiority and stopping the trial early. At analysis k, reject H_0 if:

$$Z_k \geq b_k^{\text{efficacy}} \tag{23}$$

For O'Brien-Fleming with K looks and one-sided α, boundaries are computed numerically by solving:

$$P_{H_0}\left(\bigcup_{k=1}^{K}\{Z_k \geq b_k\}\right) = \alpha \tag{24}$$

This requires integrating the multivariate normal distribution of $(Z_1, Z_2, \ldots, Z_K)$, accounting for correlation structure induced by overlapping data.

4.3.2 Translation to Effect Size Scales

Boundaries are calculated on the standardized Z-score scale but must be translated to clinically meaningful effect sizes for interpretation:

Mean Difference (Continuous Outcome):

$$\delta_k = Z_k \times \text{SE}(\delta_k) = Z_k \times \sigma\sqrt{\frac{2}{n_k}} \tag{25}$$

Hazard Ratio (Time-to-Event):

$$\log(\text{HR}_k) = Z_k \times \text{SE}(\log \text{HR}_k) = Z_k \times \sqrt{\frac{4}{d_k}} \tag{26}$$

$$\text{HR}_k = \exp\left(\frac{Z_k \times 2}{\sqrt{d_k}}\right) \tag{27}$$

where d_k is number of events at analysis k.

Odds Ratio (Binary Outcome):

$$\log(\text{OR}_k) = Z_k \times \text{SE}(\log \text{OR}_k) \tag{28}$$

4.3.3 Example: HPTN 083 Boundary Translation

The HPTN 083 HIV prevention trial (Landovitz et al., NEJM 2021) used a 4-look O'Brien-Fleming design:

Analysis	Events (d_k)	Z-boundary	HR Boundary*	Interpretation
Look 1	44 (25%)	4.333	0.39	Require 61% risk reduction
Look 2	88 (50%)	2.963	0.66	Require 34% risk reduction
Look 3	132 (75%)	2.359	0.82	Require 18% risk reduction
Look 4	176 (100%)	1.993	0.91	Require 9% risk reduction

Table 9: HPTN 083 O'Brien-Fleming boundaries

**HR boundary note:* The translation from Z-boundary to HR boundary assumes the log-rank test statistic relationship $Z = \log(\widehat{\text{HR}})/\widehat{\text{SE}}(\log(\text{HR}))$. This approximation is standard but depends on proportional hazards and the specific test statistic used. For exact boundary translation, consult with a biostatistician familiar with the trial's analysis plan.

The trial stopped at Look 1 with observed HR = 0.29, exceeding the efficacy boundary (0.39).

4.3.4 Futility Boundaries

Futility boundaries allow early termination when accumulating evidence suggests the trial is unlikely to succeed. At analysis k, stop for futility if:

$$Z_k \leq b_k^{\text{futility}} \tag{29}$$

Binding vs. Non-Binding Futility:

- **Binding futility:** Trial MUST stop if boundary crossed; Type I error calculation assumes stopping occurs
- **Non-binding futility (recommended):** Trial MAY stop, but can also continue; Type I error maintained regardless

FDA and EMA prefer **non-binding futility** to preserve trial integrity and avoid forcing premature termination based on interim data that may be misleading.

Conditional Power Futility:

A common futility rule is to stop if conditional power (probability of eventual success given interim data) falls below a threshold (e.g., 20%):

$$\text{CP}_k = P_{H_1}\left(Z_K \geq b_K^{\text{efficacy}} \mid Z_k\right) \tag{30}$$

If $\text{CP}_k < 0.20$, futility stopping may be considered.

4.4 Implementation Details

4.4.1 Sample Size Inflation Factor

Group sequential designs require slightly larger maximum sample size than fixed-sample designs to maintain power. The **inflation factor** is:

$$\text{IF} = \frac{N_{\text{GSD}}}{N_{\text{fixed}}} \tag{31}$$

Boundary Type	$K = 2$	$K = 3$	$K = 4$	$K = 5$
O'Brien-Fleming	1.01	1.02	1.02	1.03
Pocock	1.10	1.14	1.16	1.17
Linear ($\alpha \propto t$)	1.05	1.07	1.08	1.09

Table 10: Sample size inflation factors by boundary type and number of looks

Interpretation: O'Brien-Fleming designs inflate maximum sample size by only 2–3%, a small price for potential early stopping. Pocock designs pay a steeper penalty (14–17%) due to aggressive early boundaries.

4.4.2 Expected Sample Size

While maximum sample size increases slightly, **expected sample size** (average across many trials) often decreases substantially, especially under the alternative hypothesis:

$$\text{ESS}_{H_1} = \sum_{k=1}^{K} P(\text{stop at look } k \mid H_1) \times n_k \tag{32}$$

Example (4-look O'Brien-Fleming, 90% power):

- Maximum N: 1,020 (2% inflation)
- Expected N under H_1: 850 (17% reduction)
- Expected N under H_0: 950 (7% reduction)

4.5 Regulatory Considerations

4.5.1 FDA Guidance for Industry: Adaptive Designs (November 2019)

The FDA issued comprehensive guidance recognizing group sequential designs as well-established methodology:

1. **Classification:**

   "Group sequential designs are the simplest and most established type of adaptive design and are widely used in confirmatory trials."

2. **Type I Error Control:**

   "For group sequential designs, it is straightforward to strongly control the Type I error rate using established statistical methods (e.g., alpha-spending functions)."

3. **Data Monitoring Committees:**

   "For confirmatory trials, FDA recommends establishing an independent Data Monitoring Committee (DMC) to review unblinded interim data and make stopping recommendations."

4.5.2 DSMB/DMC Charter Requirements

Regulatory agencies expect detailed Data Safety Monitoring Board (DSMB) charters for trials with interim analyses:

1. **Composition:** 3–7 members with relevant expertise (biostatistics, clinical, ethics); no financial conflicts
2. **Responsibilities:** Review unblinded interim data; assess benefit-risk balance; recommend continuation, modification, or termination
3. **Meeting Schedule:** Pre-specified interim analyses per GSD design; ad hoc meetings for safety concerns
4. **Statistical Procedures:** O'Brien-Fleming boundaries or other pre-specified rules; conditional power calculations

5 Bayesian Predictive Power

5.1 Theoretical Foundation

Bayesian Predictive Probability of Success (PPoS) provides a framework for interim decision-making in clinical trials by computing the probability that a trial will succeed at its final analysis, given accumulated interim data and prior beliefs about treatment effects. Unlike frequentist conditional power (which conditions on a fixed parameter value), Bayesian predictive power integrates over the posterior distribution, properly accounting for parameter uncertainty.

5.1.1 Core Framework

Let θ denote the treatment effect parameter (e.g., log hazard ratio, log odds ratio, mean difference). Predictive probability of success is:

$$\text{PPoS} = \int P(\text{Trial Success at Final Analysis} \mid \text{Final Data}, \theta) \times \pi(\theta \mid \text{Interim Data})\, d\theta \tag{33}$$

This integrates **conditional power** over the **posterior distribution** of θ:

$$\boxed{\text{PPoS} = \int \text{CP}(\theta) \times \pi(\theta \mid D_{\text{interim}})\, d\theta} \tag{34}$$

5.1.2 Comparison: Conditional Power vs. Predictive Power

Aspect	Conditional Power (Frequentist)	Predictive Power (Bayesian)
Parameter Treatment	Fixed at specific value θ^*	Distribution $\pi(\theta\|\text{data})$
Formula	$P(\text{Success} \mid \theta = \theta^*)$	$\int P(\text{Success}\|\theta)\pi(\theta\|\text{data})d\theta$
Uncertainty	Ignores parameter uncertainty	Fully accounts for uncertainty
Interpretation	"If true effect is θ^*, prob of success"	"Given what we know now, prob of success"
Computational	Single power calculation	Integration over posterior
Prior Influence	None	Depends on prior specification

Table 11: Comparison of conditional vs. predictive power

5.1.3 Advantages of Predictive Power

1. **Natural Interpretation:** Answers the question sponsors actually care about: "What's the probability we'll succeed?"
2. **Accounts for Uncertainty:** Early in a trial with limited data, parameter estimates are highly uncertain. Predictive power appropriately reflects this uncertainty.
3. **Coherent Updating:** As data accumulate, posterior concentrates around true effect, and predictive power converges to 0 or 1.
4. **Prior Incorporation:** Allows incorporation of external information (historical data, expert opinion) through prior distribution.

5.2 Conjugate Prior Families

Conjugate priors enable analytical computation of posterior and predictive distributions, offering computational efficiency and mathematical elegance. Zetyra supports three conjugate families covering common endpoint types.

5.2.1 Beta-Binomial Model (Binary Endpoints)

Setup:

- Outcome: Response rate p (probability of success)
- Prior: $p \sim \text{Beta}(\alpha_0, \beta_0)$
- Likelihood: $X \mid p \sim \text{Binomial}(n, p)$, where X = number of successes
- Posterior: $p \mid X \sim \text{Beta}(\alpha_0 + x, \beta_0 + n - x)$

Posterior Mean:

$$\mathbb{E}[p \mid x, n] = \frac{\alpha_0 + x}{\alpha_0 + \beta_0 + n} \tag{35}$$

This is a weighted average of prior mean $\alpha_0/(\alpha_0 + \beta_0)$ and sample proportion x/n, with weights determined by prior pseudo-sample size $(\alpha_0 + \beta_0)$ vs. actual sample size n.

Posterior Predictive Distribution:

For future y successes in m additional patients:

$$P(Y = y \mid x, n) = \binom{m}{y} \frac{B(\alpha_0 + x + y, \beta_0 + n - x + m - y)}{B(\alpha_0 + x, \beta_0 + n - x)} \tag{36}$$

where $B(a, b)$ is the beta function: $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$.

This is the **Beta-Binomial distribution** with parameters $(m, \alpha_0 + x, \beta_0 + n - x)$.

Prior Specification Guidelines:

Prior	Interpretation	When to Use
Beta(1,1)	Uniform (non-informative)	No prior information; let data dominate
Beta(0.5, 0.5)	Jeffreys prior	Non-informative but emphasizes extremes
Beta(a, b), $\mathbb{E}[p] = p_0$	Skeptical prior at null	Require strong evidence to overcome null
Beta(a, b), $\mathbb{E}[p] = p_1$	Enthusiastic prior	Historical evidence supports treatment

Table 12: Prior specification guidelines for beta-binomial model

5.2.2 Normal-Normal Model (Continuous Endpoints)

Setup:

- Outcome: Mean difference μ (e.g., change from baseline)
- Prior: $\mu \sim N(\mu_0, \tau_0^2)$
- Likelihood: $\bar{X} \mid \mu \sim N(\mu, \sigma^2/n)$, where $\bar{X}$ = sample mean, σ^2 known
- Posterior: $\mu \mid \bar{X} \sim N(\mu_{\text{post}}, \tau_{\text{post}}^2)$

Precision-Weighted Posterior:

Posterior mean is a precision-weighted average of prior and data:

$$\mu_{\text{post}} = \frac{\mu_0/\tau_0^2 + n\bar{X}/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \tag{37}$$

Posterior precision (inverse variance) sums prior and data precisions:

$$\frac{1}{\tau_{\text{post}}^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \tag{38}$$

Interpretation:

- Strong prior (small τ_0^2): Posterior pulled toward μ_0
- Weak prior (large τ_0^2): Posterior dominated by data $\bar{X}$
- Large sample (large n): Data overwhelms prior

Posterior Predictive Distribution:

For future sample mean $\bar{Y}$ from m additional patients:

$$\bar{Y} \mid \bar{X} \sim N\left(\mu_{\text{post}}, \tau^2_{\text{post}} + \frac{\sigma^2}{m}\right) \tag{39}$$

Predictive variance = posterior variance (parameter uncertainty) + sampling variance (future data variability).

5.2.3 Gamma-Poisson Model (Count Endpoints)

Setup:

- Outcome: Event rate λ (events per person-time)
- Prior: $\lambda \sim \text{Gamma}(\alpha_0, \beta_0)$
- Likelihood: $X \mid \lambda \sim \text{Poisson}(n\lambda)$, where X = total events in n person-years
- Posterior: $\lambda \mid X \sim \text{Gamma}(\alpha_0 + x, \beta_0 + n)$

Posterior Mean:

$$\mathbb{E}[\lambda \mid x, n] = \frac{\alpha_0 + x}{\beta_0 + n} \tag{40}$$

5.3 Predictive Probability Calculations

5.3.1 Analytical Computation (Beta-Binomial Example)

Scenario: Single-arm Phase II trial, $N = 40$ patients

- Null: $p_0 = 0.30$ (standard therapy response rate)
- Alternative: $p_1 = 0.50$ (target response rate)
- Success criterion: $P(p > 0.30 \mid \text{Final Data}) > 0.95$
- Prior: Beta(1,1) = Uniform
- Interim: $n = 20$ patients, $x = 8$ responses (40% observed rate)

Step 1: Posterior Distribution

$$p \mid x = 8, n = 20 \sim \text{Beta}(1 + 8, 1 + 12) = \text{Beta}(9, 13) \tag{41}$$

Step 2: Posterior Probability ($p > 0.30$)

$$P(p > 0.30 \mid \text{data}) = \int_{0.30}^{1} \text{Beta}(9, 13)\, dp = 0.814 \tag{42}$$

Interpretation: 81.4% posterior probability that true response rate exceeds 30%.

Step 3: Predictive Probability

For trial to succeed, need ≥ 20 total responses out of 40 patients $\rightarrow$ need ≥ 12 responses in remaining 20 patients.

$$\text{PPoS} = P(Y \geq 12 \mid x = 8, n = 20) \tag{43}$$

where $Y \sim \text{Beta-Binomial}(m = 20, \alpha = 9, \beta = 13)$.

Compute using beta function:

$$P(Y = y \mid x, n) = \binom{20}{y} \frac{B(9 + y, 13 + 20 - y)}{B(9, 13)} \tag{44}$$

$$\text{PPoS} = \sum_{y=12}^{20} P(Y = y \mid x, n) = 0.367 \tag{45}$$

Interpretation: 36.7% probability of trial success if continued. This exceeds typical futility threshold (10–20%), so DSMB would likely recommend continuation.

5.3.2 Monte Carlo Simulation Algorithm

When analytical solutions are unavailable (e.g., non-conjugate priors, complex success criteria), Monte Carlo simulation provides general-purpose approximation.

Algorithm (Berry et al. 2010):

Input: Interim data D_{interim}, prior $\pi(\theta)$, $N_{\max}$, success criterion

Initialize: $K \leftarrow$ Number of simulations (e.g., 10,000); success_count $\leftarrow 0$

For $k = 1$ to K:

1. Draw parameter from posterior: $\theta^{(k)} \sim \pi(\theta \mid D_{\text{interim}})$
2. Simulate future data: $D_{\text{future}}^{(k)} \sim f(\text{data} \mid \theta^{(k)}, N_{\text{remaining}})$
3. Combine: $D_{\text{combined}} \leftarrow D_{\text{interim}} \cup D_{\text{future}}^{(k)}$
4. Compute final posterior: $\pi_{\text{final}}^{(k)} \leftarrow \pi(\theta \mid D_{\text{combined}})$
5. **If** success_criterion$(\pi_{\text{final}}^{(k)})$: success_count $\leftarrow$ success_count $+1$

Return: PPoS $\leftarrow$ success_count$/K$

Convergence: Standard error of PPoS estimate $\approx \sqrt{\text{PPoS}(1 - \text{PPoS})/K}$. For PPoS ≈ 0.5, $K = 10,000$ yields SE ≈ 0.005 (acceptable precision).

5.4 Decision Framework

5.4.1 Phase II Go/No-Go Decision Thresholds

PPoS Range	Recommendation	Rationale
$< 10\%$	Stop for futility	$< 10\%$ chance of Phase III success
10–30%	Borderline; re-evaluate	Consider design modifications, biomarker refinement
30–50%	Continue with caution	May proceed if unmet medical need high
$> 50\%$	Proceed to Phase III	$> 50\%$ success probability justifies investment
$> 85\%$	High confidence; early graduation	Very promising; I-SPY 2 uses 85% threshold

Table 13: Phase II Go/No-Go decision thresholds

5.4.2 Integration with Business Metrics

$$\text{Expected Value} = P(\text{Success}) \times \text{NPV}(\text{Success}) - P(\text{Failure}) \times \text{Cost}(\text{Failure}) \tag{46}$$

Where:

- $P(\text{Success})$ = PPoS from Bayesian calculation
- NPV(Success) = \$500M–\$2B (blockbuster drug)
- $P(\text{Failure}) = 1 - \text{PPoS}$
- Cost(Failure) = \$50–100M (wasted Phase III investment)

Example:

- PPoS = 0.40 (40% success probability)
- NPV(Success) = \$800M
- Cost(Failure) = \$80M

Expected Value = $0.40 \times \$800\text{M} - 0.60 \times \$80\text{M} = \$320\text{M} - \$48\text{M} = \$\mathbf{272}\text{M} \rightarrow$ Proceed to Phase III

5.5 Regulatory Considerations

5.5.1 FDA Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials (February 2010)

This foundational guidance established FDA's openness to Bayesian methods:

1. **Acceptance of Bayesian Designs:**

   > "FDA has no fundamental objection to the use of Bayesian methods in medical device trials, provided the design has appropriate operating characteristics and is adequately justified."

2. **Operating Characteristics:**

   > "Bayesian designs must demonstrate adequate frequentist operating characteristics (Type I error control, power) through simulation studies covering a range of true parameter values."

5.5.2 FDA Draft Guidance: Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products (January 12, 2026)

Note: This is *draft* guidance, not yet finalized for implementation. Sponsors should monitor FDA's website for the final guidance and consult with their review division before incorporating Bayesian methods into regulatory submissions.

In a significant policy expansion, FDA released draft guidance extending Bayesian acceptance to drugs and biologics (satisfying PDUFA VII commitment Section I.L.4.f):

1. **Commissioner Statement:**

   > "Bayesian methodologies help address two of the biggest problems of drug development: high costs and long timelines. Providing clarity around modern statistical methods will help sponsors bring more cures and meaningful treatments to patients faster and more affordably."
   > — **FDA Commissioner Marty Makary, M.D., M.P.H.** (January 12, 2026)

2. **Applications Endorsed:**

   - **Futility/Success Determination:** Predictive probability for interim go/no-go decisions
   - **Dose Selection:** Posterior distributions for optimal dose identification
   - **Subgroup Identification:** Borrowing strength across subgroups via hierarchical models
   - **Incorporating Historical Data:** Informative priors from prior studies, real-world evidence
   - **Rare & Pediatric Diseases:** Leveraging external evidence when patient populations are smaller

6 Validation Framework

6.1 Overview and Methodology

Zetyra calculators undergo comprehensive external validation through three complementary approaches: (1) software benchmarking against established reference implementations, (2) analytical formula verification using closed-form solutions, and (3) published clinical trial replication. This multi-faceted validation strategy ensures accuracy, identifies implementation errors, and demonstrates real-world applicability.

Validation Principles:

All validation code, test data, and results are publicly available at `github.com/evidenceinthewild/zetyra-validation` under MIT license. This transparency enables:

- **Independent verification:** Anyone can reproduce our validation results
- **Continuous validation:** GitHub Actions automatically runs 51 tests on every code change
- **Regulatory scrutiny:** Auditors can examine validation methodology and results
- **Community contribution:** External statisticians can propose additional test cases

How to Validate Us Yourself

Reproduce our validation in 5 minutes:

1. Clone the repository: `git clone https://github.com/evidenceinthewild/zetyra-validation`
2. Install dependencies: `pip install -r requirements.txt`
3. Run the test suite: `pytest tests/ -v`
4. Compare results to Table 15 (51 tests, all should pass)
5. Examine individual test cases in `tests/` to understand validation methodology

For regulatory submissions: The validation report (`docs/validation_report.pdf`) provides detailed test specifications, expected vs. actual results, and traceability to requirements.

6.2 Acceptance Criteria

For each validation test, we pre-specified acceptance criteria before implementation:

Calculator	Metric	Tolerance	Rationale
GSD	Z-score boundary deviation	± 0.05	Standard tolerance for commercial software
CUPED	Variance reduction factor	Exact match	Analytical formula has no numerical error
Bayesian	Predictive probability	± 0.001	Rounding error in numerical integration

Table 14: Pre-specified acceptance criteria

6.3 Validation Summary

Calculator	Tests	Passed	Max Deviation	Reference
GSD	30	30 (100%)	0.0046 z-score	gsDesign R package
CUPED	12	12 (100%)	Exact match	Analytical VRF $= 1 - \rho^2$
Bayesian	9	9 (100%)	Exact match	Conjugate prior formulas
Total	**51**	**51 (100%)**	**0.0046**	Multiple benchmarks

Table 15: Validation results summary

All tests passed with maximum deviation **10-fold smaller** than pre-specified tolerance (0.0046 vs. 0.05 for GSD). This level of numerical agreement is consistent with expectations for software used to support regulatory submissions; sponsors should confirm fit-for-purpose within their quality system.

6.4 Group Sequential Design Validation

Group sequential design calculations are validated against the **gsDesign R package** (version 3.6.4), widely recognized as the gold standard for sequential trial design. gsDesign is maintained by Keaven Anderson at Merck and is cited in thousands of FDA submissions.

6.4.1 gsDesign Benchmark Results

Design	Looks	Boundaries Tested	Max Deviation	Mean Deviation	Status
OF_2	2	2	0.0000	0.0000	✓Pass
OF_3	3	3	0.0001	0.0001	✓Pass
OF_4	4	4	0.0017	0.0009	✓Pass
OF_5	5	5	0.0046	0.0023	✓Pass
Pocock_2	2	2	0.0000	0.0000	✓Pass
Pocock_3	3	3	0.0002	0.0001	✓Pass
Pocock_4	4	4	0.0008	0.0004	✓Pass

Table 16: GSD validation against gsDesign R package (23 boundary comparisons)

Interpretation:

- **All 23 boundaries passed** with deviations well below 0.05 tolerance
- **Maximum deviation (0.0046)** occurred in OF_5 design (5 looks, most complex scenario)
- **Mean absolute deviation (0.0008)** demonstrates consistent accuracy across designs
- **Accuracy improves with fewer looks** (OF_2, OF_3 exact match to gsDesign)

6.4.2 Published Trial Replication

HPTN 083 Trial (HIV Prevention, 2021)

The HIV Prevention Trials Network (HPTN) 083 trial compared long-acting injectable cabotegravir to daily oral TDF/FTC for HIV pre-exposure prophylaxis in 4,566 cisgender men and transgender women (Landovitz et al., NEJM 2021;385:595-608). The trial used a **4-look O'Brien-Fleming design** with information fractions [0.25, 0.50, 0.75, 1.00] based on events (HIV infections).

Analysis	Info %	gsDesign Z	Zetyra Z	Deviation
Look 1	25%	4.049	4.0444	0.0046
Look 2	50%	2.863	2.8598	0.0032
Look 3	75%	2.337	2.3351	0.0019
Look 4	100%	2.024	2.0222	0.0018

Table 17: HPTN 083 boundary replication

Trial Outcome: The trial stopped early at Look 1 (44 events, 26% information) with observed HR = 0.29 (95% CI: 0.14–0.58), crossing the efficacy boundary and demonstrating superiority of cabotegravir.

6.5 CUPED Variance Reduction Validation

CUPED variance reduction calculations are validated against **analytical formulas** with closed-form solutions. The variance reduction factor has an exact mathematical expression:

$$\text{VRF} = 1 - \rho^2 \tag{47}$$

This enables exact validation: any numerical implementation should reproduce VRF $= 1 - \rho^2$ to machine precision.

Correlation (ρ)	Analytical VRF	Zetyra VRF	Deviation	SS Reduction
0.0	1.0000	1.0000	0.0000	0%
0.3	0.9100	0.9100	0.0000	9%
0.5	0.7500	0.7500	0.0000	25%
0.6	0.6400	0.6400	0.0000	36%
0.7	0.5100	0.5100	0.0000	49%
0.9	0.1900	0.1900	0.0000	81%

Table 18: CUPED variance reduction factor validation

Interpretation: Exact match to analytical formula across all correlation values. No numerical error detected; deviations are zero to machine precision.

6.6 Bayesian Predictive Power Validation

Bayesian predictive power calculations are validated against **analytical solutions** for conjugate prior families.

6.6.1 Beta-Binomial Validation

Test Scenario: Phase II single-arm trial, $N = 40$ patients

- Success criterion: ≥ 20 responses
- Prior: Beta(1,1) – uniform
- Interim: $x = 8$ responses in $n = 20$ patients

Results:

- **Posterior probability:** Zetyra = 0.814, Analytical = 0.814 (exact match) ✓
- **Predictive probability:** Zetyra = 0.367, Analytical = 0.367 (exact match) ✓

6.6.2 Normal-Normal Validation

Test Scenario: Two-arm trial, $N = 100$ per arm

- Effect observed: $\delta_{\text{obs}} = 0.3$ SD at interim ($n = 50$ per arm)
- Prior: $N(0.5, 1000)$ – essentially flat

Results:

- **Posterior mean:** Zetyra = 0.300, Analytical = 0.300 ✓
- **Posterior SE:** Zetyra = 0.141, Analytical = 0.141 ✓
- **Predictive power:** Zetyra = 0.52, Analytical = 0.52 ✓

6.7 Validation Conclusions

Zetyra's three statistical calculators passed **51 automated validation tests** (30 GSD, 12 CUPED, 9 Bayesian) with 100% success rate. Key achievements:

- **Accuracy:** Maximum deviation 0.0046 z-score, **10× better** than industry standard (±0.05 tolerance)
- **Transparency:** All validation code public (MIT license)
- **Benchmarks:** Matches gsDesign R package (gold standard), analytical formulas, and published trial designs
- **Regulatory Readiness:** Methodology aligned with FDA software validation guidance

Competitive Advantage: Commercial alternatives (East, PASS, ADDPLAN) do not typically provide public, independently reproducible validation suites. Zetyra's open-source validation suite enables independent verification with quantified accuracy metrics (0.0046 max deviation).

7 Case Studies

This section presents four detailed case studies demonstrating real-world applications of Zetyra's calculators. Each case study quantifies cost savings, timeline acceleration, and statistical efficiency gains achievable through modern trial design methods.

Important: Illustrative Scenarios

These case studies are illustrative scenarios, not retrospective analyses of actual trials. They are constructed from:

- Published trial parameters (HPTN 083, HeartMate II, industry benchmarks)
- Literature-supported assumptions (Walters et al., 2019 correlation data; Jennison & Turnbull, 2000 GSD efficiency; Berry et al., 2010 Bayesian frameworks)
- Standard industry cost estimates ($40K–$60K per patient for Phase III)

Actual benefits depend heavily on: trial-specific characteristics, endpoint correlation strength, interim timing, regulatory feedback, and execution quality. ROI calculations separate "trial cost savings" (more defensible) from "revenue timing gains" (highly assumption-dependent). See Section 7.5 for sensitivity analysis.

7.1 Case Study 1: Oncology Phase II Trial (CUPED)

Trial Context: A mid-sized biotech company is developing a novel antibody-drug conjugate for HER2-positive metastatic breast cancer. They plan a single-arm Phase II trial to evaluate objective response rate (ORR) before advancing to Phase III.

Standard Design (No CUPED):

- Null hypothesis: $p_0 = 0.35$ (historical standard therapy ORR)
- Alternative hypothesis: $p_1 = 0.55$ (target ORR for novel agent)
- Significance: $\alpha = 0.025$ (one-sided), Power: 90%
- Required N: 240 patients
- Per-patient cost: \$50,000
- Total trial cost: \$12.0 million

CUPED-Adjusted Design:

Using baseline tumor burden correlation $\rho = 0.55$ (from Walters et al., 2019):

$$\text{VRF} = 1 - 0.55^2 = 0.6975 \tag{48}$$

$$N_{\text{CUPED}} = 240 \times 0.6975 = 167.4 \approx \mathbf{168} \text{ patients} \tag{49}$$

Sample size reduction: $240 - 168 = \mathbf{72}$ **patients (30%)**

Impact Quantification:

Metric	Standard Design	CUPED Design	Savings
Sample Size	240 patients	168 patients	72 patients (30%)
Enrollment Duration	12 months	8.4 months	3.6 months
Drug Cost	\$9.6M	\$6.7M	\$2.9M
Site Management	\$1.2M	\$0.84M	\$0.36M
Monitoring/CRO	\$1.2M	\$0.84M	\$0.36M
Total Cost	**\$12.0M**	**\$8.4M**	**\$3.6M (30%)**

Table 19: Case Study 1: CUPED impact on Phase II oncology trial

7.2 Case Study 2: Cardiovascular Phase III Trial (GSD)

Trial Context: A large pharmaceutical company is developing a novel PCSK9 inhibitor for secondary prevention of major adverse cardiovascular events (MACE). Primary endpoint: time to first MACE.

Standard Fixed-Sample Design:

- Target hazard ratio: HR = 0.75 (25% relative risk reduction)
- Required events: $D_{\text{fixed}} = 430$ events
- Total duration: 48 months
- Total cost: \$76.8 million

Group Sequential Design (4-Look O'Brien-Fleming):

Analysis	Events	Month	Z-boundary	HR Boundary
Interim 1	108 (25%)	30	3.897	0.55
Interim 2	215 (50%)	36	2.754	0.70
Interim 3	323 (75%)	42	2.250	0.80
Final	430 (100%)	48	2.014	0.86

Table 20: Case Study 2: GSD boundaries for cardiovascular trial

Actual Trial Outcome:

Interim Analysis 2 (Month 36, 220 events observed):

- Observed HR: 0.68 (95% CI: 0.54–0.86)
- Z-statistic: 2.89
- Boundary: $Z = 2.754$
- **Decision: STOP FOR EFFICACY** ($Z = 2.89 > 2.754$)

Impact Quantification:

Metric	Fixed Design	GSD (Stopped Early)	Savings
Total Duration	48 months	36 months	**12 months (25%)**
Events Required	430	220	210 events
Monitoring Cost	\$76.8M	\$58.7M	**\$18.1M (24%)**
Time to Submission	Month 54	Month 42	**12 months earlier**

Table 21: Case Study 2: GSD impact on cardiovascular Phase III trial

Revenue Impact:

- Peak annual sales (projected): \$3.5 billion
- Additional revenue from 12-month earlier approval: \$2.9–3.5 billion
- Net present value (NPV) gain: \$2.4 billion (discounted at 10%)

7.3 Case Study 3: Rare Disease Trial (Bayesian)

Trial Context: A small biotech company is developing a gene therapy for Duchenne muscular dystrophy (DMD). With only 200–300 eligible patients in the U.S., traditional Phase II/III paradigm is infeasible. Maximum enrollment: $N = 30$ patients.

Trial Design:

- Primary endpoint: Mean change in 6-minute walk distance (6MWD) at 12 months
- Success criterion: $P(\Delta 6\text{MWD} > 30\text{m} \mid \text{data}) > 0.90$
- Interim analysis: After $n = 20$ patients
- Prior: $\mu_0 = 0$, $\tau_0 = 40$ (skeptical)

Interim Analysis ($n = 20$):

- Observed: $\bar{\Delta}$6MWD = +35 meters, SD = 48 meters

Posterior Calculation:

$$\frac{1}{\tau^2_{\text{post}}} = \frac{1}{40^2} + \frac{20}{48^2} = 0.000625 + 0.008681 = 0.009306 \tag{50}$$

$$\tau_{\text{post}} = 10.37 \text{ meters} \tag{51}$$

$$\mu_{\text{post}} = \frac{0/40^2 + 20 \times 35/48^2}{0.009306} = 32.6 \text{ meters} \tag{52}$$

Posterior Probability:

$$P(\Delta\text{6MWD} > 30 \mid \text{data}) = P\left(Z > \frac{30 - 32.6}{10.37}\right) = P(Z > -0.25) = 0.60 \tag{53}$$

Predictive Probability: Monte Carlo simulation (10,000 iterations): PPoS = **0.42** (42% probability of success if continued)

Decision Framework:

PPoS	Interpretation	Decision
< 0.10	Futility	Stop, no-go
0.10–0.30	Low probability	Likely no-go
0.30–0.70	Moderate probability	Continue enrollment
> 0.70	High probability	Continue with confidence

With PPoS = 0.42, decision: **CONTINUE ENROLLMENT**

Final Analysis ($N = 30$):

- Observed: $\bar{\Delta}$6MWD = +38 meters
- Final posterior: $P(\Delta\text{6MWD} > 30 \mid \text{final data}) = 0.78$
- Did not meet pre-specified 0.90 Bayesian threshold
- **Company decision: GO to Pivotal Trial**

Decision rationale: Although the 0.90 posterior probability threshold was not met, the company proceeded based on: (1) clinically meaningful point estimate (+38m exceeds MCID of 30m), (2) 78% posterior probability still indicates strong evidence of benefit, (3) FDA Type B meeting feedback indicated willingness to consider accelerated approval pathway given unmet need in DMD, and (4) natural history data showing progressive decline without treatment. This illustrates that Bayesian thresholds inform but do not dictate decisions—regulatory, clinical, and commercial context also matter.

7.4 Case Study 4: Comparative Program Analysis

Development Program: Non-small cell lung cancer (NSCLC) immunotherapy combination

Traditional vs. Zetyra-Optimized Development:

Metric	Traditional	Zetyra	Savings
Phase II Sample Size	80	53 (avg: 48)	32 patients (40%)
Phase II Duration	12 months	8 months	4 months (33%)
Phase II Cost	\$4.0M	\$2.4M	\$1.6M (40%)
Phase III Sample Size	500	515 max (425 expected)	75 expected (15%)
Phase III Duration	54 months	42 months (expected)	12 months (22%)
Phase III Cost	\$100M	\$87.5M	\$12.5M (13%)
Total Duration	**66 months**	**50 months**	**16 months (24%)**
Total Cost	**\$104M**	**\$89.9M**	**\$14.1M (14%)**
Time to BLA	Month 72	Month 56	16 months earlier

Table 22: Case Study 4: Traditional vs. Zetyra-optimized development program

Revenue Impact:

Assuming \$2B peak annual sales with 10-year exclusivity:

- 16 months earlier launch
- Additional revenue: \$2B $\times$ (16/12) = \$**2.67**B
- NPV (discounted at 10%): \$2.1B

Return on Investment (ROI):

Trial Cost Savings ROI (more defensible):

$$\text{ROI}_{\text{cost}} = \frac{\$14.1\text{M savings} - \$0.15\text{M implementation}}{\$0.15\text{M}} = \mathbf{93}\times \tag{54}$$

Note: Revenue timing gains (\$2.1B NPV) are highly assumption-dependent and shown separately. See Section 7.5 for sensitivity analysis.

7.5 ROI Sensitivity Analysis

The ROI estimates above depend on multiple assumptions. This section provides sensitivity analysis to help readers assess applicability to their specific context.

Trial Cost Savings Sensitivity (Case Study 4):

Scenario	Assumptions	Cost Savings	ROI
Conservative	$\rho = 0.35$, 1 interim, 50% GSD benefit	\$7.0M	46$\times$
Base Case	$\rho = 0.50$, 2 interims, 75% GSD benefit	\$14.1M	93$\times$
Optimistic	$\rho = 0.60$, 3 interims, 100% GSD benefit	\$21.0M	139$\times$

Table 23: Trial cost savings sensitivity (excludes revenue timing)

Revenue Timing Sensitivity (illustrative only):

Peak Sales	Time Gained	Discount Rate	P(Success)	Risk-Adj NPV
\$500M	12 months	12%	60%	\$267M
\$1B	16 months	10%	50%	\$667M
\$2B (base)	16 months	10%	100%	\$2.1B
\$2B	16 months	10%	50%	\$1.05B

Table 24: Revenue timing NPV sensitivity (highly assumption-dependent)

Key Takeaway

Trial cost savings (reduced patient enrollment, shorter timelines) represent the most defensible ROI component and alone justify Zetyra adoption at 46–139× ROI depending on trial characteristics.
Revenue timing gains can be substantial but depend heavily on peak sales projections, probability of technical/regulatory success, and competitive dynamics. These should be evaluated within each sponsor's specific commercial forecasting framework.

8 Conclusions

8.1 Summary of Capabilities

Zetyra provides a validated, integrated platform of three statistical calculators addressing complementary inefficiencies in clinical trial design:

CUPED (Covariate-Adjusted Power Analysis):

- Leverages baseline-outcome correlations to reduce sample size by 15–35%
- Validated against analytical variance reduction formula (VRF $= 1-\rho^2$) with exact matches
- Methodology aligned with FDA May 2023 guidance encouraging covariate adjustment (note: FDA guidance supports the statistical practice of covariate adjustment; CUPED is Zetyra's implementation)

Group Sequential Design:

- Enables interim efficacy/futility monitoring with rigorous Type I error control
- Validated against gsDesign R package (gold standard) with maximum deviation 0.0046 z-score
- O'Brien-Fleming boundaries require only 2–3% sample size inflation while enabling 15–40% expected sample size reduction

Bayesian Predictive Power:

- Computes probability of trial success given interim data and prior beliefs
- Validated against analytical conjugate prior formulas (exact matches)
- Enables quantitative go/no-go decisions vs. binary p-value thresholds

8.2 Competitive Advantages

Capability	Zetyra	East	PASS	ADDPLAN	nQuery
CUPED Calculator	✓	–	–	–	–
Group Sequential Design	✓	✓	✓	✓	–
Bayesian Predictive Power	✓	–	–	✓	–
All Three Integrated	✓	–	–	–	–
Public Validation Suite	✓	–	–	–	–
Web-Based (no install)	✓	–	–	–	–
Monthly Subscription	✓	–	–	–	–
Annual Cost*	**$1,188**	$15,000	$8,000	$12,000	$6,000

Table 25: Competitive comparison of clinical trial design software (as of January 2026)

**Pricing based on published list prices as of January 2026. Actual pricing may vary by negotiation, academic discounts, or bundling. Competitor capabilities assessed from publicly available product documentation; vendors may offer additional features not reflected here.*

8.3 Case Study ROI Summary

Case Study	Cost Savings	Time Savings	ROI
Oncology Phase II (CUPED)	$3.6M (30%)	3.6 months	10,000×
CV Phase III (GSD)	$18.1M (24%)	12 months	120,000×
Rare Disease (Bayesian)	Avoided $10M futile	18 months	N/A
Full Program (Integrated)	$14.1M (14%)	16 months (24%)	93,000×

Table 26: Summary of case study ROI

8.4 Conclusion

Zetyra provides a validated, integrated platform of three statistical calculators addressing complementary inefficiencies in clinical trial design. By combining rigorous methodology, transparent validation, and regulatory expertise, Zetyra enables biostatisticians to design more efficient trials without sacrificing statistical rigor or regulatory acceptability.

The case studies demonstrate that even modest adoption—a single Phase II/III program—can generate substantial cost and time savings. As regulatory agencies increasingly encourage efficient designs (FDA's 2023 CUPED guidance, 2026 Bayesian guidance published January 12, 2026), methodologies like covariate adjustment, group sequential monitoring, and Bayesian predictive power will transition from competitive advantage to industry standard.

The future of clinical trial design is transparent, validated, accessible, and efficient.

9 References

Statistical Methodology

1. Deng A, Xu Y, Kohavi R, Walker T. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. *WSDM* 2013:123-132.

2. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine* 1992;11(13):1685-1704.

3. Teerenstra S, Eldridge S, Graff M, de Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Statistics in Medicine* 2012;31(20):2169-2178.

4. Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow up measurements. *BMJ* 2001;323(7321):1123-1124.

5. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977;64(2):191-199.

6. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979;35(3):549-556.

7. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983;70(3):659-663.

8. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall/CRC Press, 2000.

9. Hwang IK, Shih WJ, De Cani JS. Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine* 1990;9(12):1439-1445.

10. Kim K, DeMets DL. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 1987;74(1):149-154.

11. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials* 1986;7(1):8-17.

12. Berry SM, Carlin BP, Lee JJ, Muller P. *Bayesian Adaptive Methods for Clinical Trials.* CRC Press, 2010.

13. Berry DA, Ho CH. One-sided sequential stopping boundaries for clinical trials: A decision-theoretic approach. *Biometrics* 1988;44(1):219-227.

14. Saville BR, Connor JT, Ayers GD, Alvarez J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clinical Trials* 2014;11(4):485-493.

15. Walters SJ, et al. Sample size estimation for randomised controlled trials with repeated assessment of patient-reported outcomes: what correlation between baseline and follow-up outcomes should we assume? *Trials* 2019;20:566.

Regulatory Guidance Documents

16. U.S. Food and Drug Administration. Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products: Guidance for Industry. May 2023.

17. U.S. Food and Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry. November 2019.

18. U.S. Food and Drug Administration. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. February 2010.

19. U.S. Food and Drug Administration. Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products: Draft Guidance for Industry. January 12, 2026.

20. European Medicines Agency. Guideline on Adjustment for Baseline Covariates in Clinical Trials. EMA/CHMP/295050/2013. February 2015.

21. European Medicines Agency. Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design. CHMP/EWP/2459/02. October 2007.

22. International Council for Harmonisation (ICH). E9 Statistical Principles for Clinical Trials. February 1998.

23. International Council for Harmonisation (ICH). E9(R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials. November 2019.

Published Clinical Trials

24. Landovitz RJ, et al. Cabotegravir for HIV prevention in cisgender men and transgender women. *NEJM* 2021;385(7):595-608.

25. Slaughter MS, et al. Advanced heart failure treated with continuous-flow left ventricular assist device. *NEJM* 2009;361(23):2241-2251.

26. Barker AD, et al. I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics* 2009;86(1):97-100.

27. Maude SL, et al. Tisagenlecleucel in children and young adults with B-cell lymphoblastic leukemia. *NEJM* 2018;378(5):439-448.

Software and Validation

28. Anderson K. gsDesign: Group Sequential Design. R package version 3.6.4. 2024. https://keaven.github.io/gsDesign/

Cost-Effectiveness and Trial Economics

29. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics* 2016;47:20-33.

30. Moore TJ, et al. Estimated costs of pivotal trials for novel therapeutic agents approved by the US Food and Drug Administration, 2015-2016. *JAMA Internal Medicine* 2018;178(11):1451-1457.

31. Sertkaya A, et al. Key cost drivers of pharmaceutical clinical trials in the United States. *Clinical Trials* 2016;13(2):117-126.

A API Documentation

Zetyra provides a RESTful API for programmatic access to all calculators.

Operational Details

- **Authentication:** API key required in request header (`X-API-Key: <your-key>`). Keys issued upon subscription.
- **Rate Limits:** 100 requests/minute (Evidence Pro), 500 requests/minute (Evidence Collective), custom limits for Enterprise.
- **Versioning:** API version in URL path (`/api/v1/`). Breaking changes require major version increment with 6-month deprecation notice.
- **Error Format:** HTTP status codes with JSON error body: `{"error": "message", "code": "ERROR_CODE"}`
- **Data Handling:** Input parameters logged for debugging (30-day retention). No PHI fields accepted or stored. SOC 2 Type II compliance in progress.
- **Availability:** 99.9% uptime SLA (Enterprise tier). Status page: `status.zetyra.com`

Base URL

```
https://zetyra-backend-394439308230.us-central1.run.app/api/v1/
```

CUPED Calculator

```
POST /cuped
Content-Type: application/json

{
  "standard_sample_size": 240,
  "correlation": 0.55
}

Response:
{
  "variance_reduction_factor": 0.6975,
  "cuped_sample_size": 168,
  "sample_size_reduction": 72,
  "percent_reduction": 30.0
}
```

Group Sequential Design

```
POST /gsd
Content-Type: application/json

{
  "k": 4,
  "alpha": 0.025,
  "beta": 0.10,
```

```
  "boundary_type": "obrien_fleming",
  "effect_size": 0.5,
  "test_type": "two_sample"
}

Response:
{
  "max_sample_size": 842,
  "efficacy_boundaries": [3.897, 2.754, 2.250, 2.014],
  "information_fractions": [0.25, 0.50, 0.75, 1.00],
  "expected_sample_size_h1": 695,
  "inflation_factor": 1.022
}
```

Bayesian Predictive Power (Binary)

```
POST /bayesian/binary
Content-Type: application/json

{
  "prior_alpha": 1,
  "prior_beta": 1,
  "observed_successes": 8,
  "observed_n": 20,
  "future_n": 20,
  "threshold": 0.30,
  "success_criterion": 0.95
}

Response:
{
  "posterior_alpha": 9,
  "posterior_beta": 13,
  "posterior_mean": 0.409,
  "posterior_prob_above_threshold": 0.814,
  "predictive_probability": 0.367,
  "recommendation": "continue"
}
```

B Notation Summary

Symbol	Definition
α	Type I error rate (typically 0.025 or 0.05)
β	Type II error rate; Power $= 1 - \beta$
ρ	Pearson correlation between baseline covariate and outcome
VRF	Variance Reduction Factor $= 1 - \rho^2$
Z_k	Test statistic at interim analysis k
b_k	Boundary value at interim analysis k
τ_k	Information fraction at analysis k
$\alpha(\tau)$	Alpha-spending function
PPoS	Predictive Probability of Success
CP	Conditional Power
HR	Hazard Ratio
$\pi(\theta)$	Prior distribution for parameter θ
$\pi(\theta\|\text{data})$	Posterior distribution for parameter θ
$B(a, b)$	Beta function $= \Gamma(a)\Gamma(b)/\Gamma(a+b)$

Table 27: Summary of mathematical notation

C Key Formulas

CUPED

$$Y_{\text{CUPED}} = Y - \theta(X - \mathbb{E}[X]) \quad \text{(CUPED estimator)}$$

$$\theta^* = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad \text{(Optimal coefficient)}$$

$$\text{VRF} = 1 - \rho^2 \quad \text{(Variance reduction)}$$

$$n_{\text{CUPED}} = n_{\text{standard}} \times (1 - \rho^2) \quad \text{(Adjusted sample size)}$$

Group Sequential Design

$$\tau_k = \frac{I_k}{I_K} \quad \text{(Information time)}$$

$$\Delta\alpha_k = \alpha(\tau_k) - \alpha(\tau_{k-1}) \quad \text{(Incremental alpha)}$$

$$\alpha_{\text{OF}}(t) = 2\left[1 - \Phi\left(\frac{Z_{\alpha/2}}{\sqrt{t}}\right)\right] \quad \text{(O'Brien-Fleming)}$$

$$\alpha_{\text{HSD}}(t; \gamma) = \alpha \times \frac{1 - e^{-\gamma t}}{1 - e^{-\gamma}} \quad \text{(Hwang-Shih-DeCani)}$$

$$\text{IF} = \frac{N_{\text{GSD}}}{N_{\text{fixed}}} \quad \text{(Inflation factor)}$$

Bayesian

$$\text{PPoS} = \int \text{CP}(\theta) \times \pi(\theta | D_{\text{interim}})\, d\theta \qquad \text{(Predictive probability)}$$

$$\mathbb{E}[p|x,n] = \frac{\alpha_0 + x}{\alpha_0 + \beta_0 + n} \qquad \text{(Beta-binomial posterior mean)}$$

$$\mu_{\text{post}} = \frac{\mu_0/\tau_0^2 + n\bar{X}/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \qquad \text{(Normal posterior mean)}$$

$$\frac{1}{\tau^2_{\text{post}}} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \qquad \text{(Posterior precision)}$$

D Glossary of Statistical Terms

Alpha (α)
Type I error rate; probability of rejecting true null hypothesis (typically 0.05 or 0.025)

Alpha-Spending Function
Function $\alpha(t)$ determining how Type I error is allocated across interim analyses in group sequential designs

ANCOVA
Analysis of Covariance; adjusting treatment comparison for baseline covariates

Bayesian Predictive Probability
Probability that trial will succeed at final analysis given interim data and prior beliefs

Boundary (Efficacy/Futility)
Threshold for stopping group sequential trial early for efficacy or futility

Conditional Power
Probability of trial success given interim data, conditioning on specific treatment effect value

Conjugate Prior
Prior distribution that, when combined with likelihood, yields posterior in same family

CUPED
Controlled-experiment Using Pre-Experiment Data; variance reduction technique using baseline measurements

DSMB/DMC
Data Safety Monitoring Board / Data Monitoring Committee; independent committee reviewing unblinded interim data

Group Sequential Design
Trial design with pre-planned interim analyses and stopping rules

Hazard Ratio (HR)
Ratio of event rates in treatment vs. control (time-to-event endpoints)

Information Fraction
Proportion of maximum statistical information available at interim analysis

O'Brien-Fleming
: Conservative alpha-spending approach with high early boundaries

Pocock
: Aggressive alpha-spending with constant boundaries across all looks

Posterior Distribution
: Updated probability distribution for parameter after observing data (Bayesian)

Prior Distribution
: Initial probability distribution for parameter before observing data (Bayesian)

Variance Reduction Factor (VRF)
: Proportional decrease in outcome variance from covariate adjustment (VRF $= 1 - \rho^2$)

Suggested Citation:

Qian L. Zetyra: A Validated Suite of Statistical Calculators for Efficient Clinical Trial Design. Technical White Paper. Evidence in the Wild, January 2026. Available at: https://zetyra.com/whitepaper

Contact:
Lu (Maggie) Qian, MS
maggie@zetyra.com
https://zetyra.com