PharmaSUG China 2026 - Paper AI-137

An Integrated, AI-Augmented R Shiny Platform for Oncology Clinical Trial Decision Intelligence -- From CDISC Data to Go/No-Go in One Unified Hub

Zhiping Yan, Dizal Pharmaceutical Co., Ltd.

ABSTRACT

Modern oncology drug development necessitates that cross-functional teams integrate safety, efficacy, pharmacokinetic/pharmacodynamic (PK/PD), and competitive benchmark evidence to make coherent and timely decisions, often within strict clinical timelines. Nevertheless, in most organizations, these analyses remain fragmented across isolated tools, manual scripts, and static reports, resulting in both operational inefficiencies and decision delays.

We introduce the Oncology Clinical Trial Data-Driven Decision-Making System, an open-architecture R Shiny application consisting of over 40 modular analytical domains. The platform takes in CDISC-standard datasets (ADSL, ADAE, ADRS, ADTR, ADTTE) and user-uploaded PK exposure or benchmark data, and then unifies them via a reactive data layer that supplies: (1) a Bayesian decision hub that implements Beta-Binomial Go/No-Go analysis, Bayesian Logistic Regression CRM with Escalation With Overdose Control (EWOC), BOIN/mTPI dose-escalation simulation, Simon two-stage design, and joint efficacy–toxicity Phase II design; (2) an AI Signal Intelligence Engine that actively scans six evidence domains and ranks actionable alerts in real-time; (3) a machine-learning safety strategy module that combines logistic regression patient risk scoring, K-means AE clustering, and sequential CUSUM monitoring; (4) a competitive benchmark intelligence engine with Meta-Analytic Predictive (MAP) Prior elicitation; and (5) an AE co-occurrence knowledge graph.

The architecture shows that modern statistical programming practices, such as reactive modular design, Bayesian computation, and lightweight ML, can transform a passive reporting tool into an active decision co-pilot. This paper details the system design, statistical methodology, implementation choices, and lessons learned, aiming to provide a reproducible blueprint for the broader biopharmaceutical programming community.

Keywords: R Shiny, Bayesian adaptive design, dose escalation, CDISC, machine learning, clinical decision support, BOIN, CRM, CUSUM, knowledge graph, DDDM, signal detection, benefit-risk, CDE PFDD

INTRODUCTION

Three structural challenges converge to produce a decision gap in oncology drug development.

First, the extreme trade-off between benefit and risk. Efficacy and toxicity often share the same mechanism of action, making dose selection irreducible to a single p-value. The 'scissors gap' — diminishing efficacy returns against steeply rising toxicity — forces decision-makers to balance survival, safety, quality of life, and treatment duration simultaneously.

Second, time pressure from multiple directions. Disease progresses relentlessly in patients, regulators demand accelerated reviews, and investors watch the clock as patent life dwindles. In this environment, every wrong Go decision can represent over $100 million in sunk cost.

Third, an explosion of data dimensions. A single Phase I trial now generates dozens of safety endpoints, multiple biomarkers, dense PK profiles, and imaging features, far exceeding what manual table review can synthesise. Approximately 45% of Phase III failures had warning signs visible in Phase I/II data, but these signals often go unrecognized until it is too late.

The Data-Driven Decision-Making (DDDM) framework bridges this gap through three paradigm shifts that collectively transform the clinical team's relationship with their data. (1) From static reports to real-time interactive dashboards: reducing signal detection latency from 4-8 weeks to same-day alerts, with

interactive Shiny dashboards replacing the traditional SAS-to-PDF-to-email workflow. (2) From fragmented single-domain analysis to a panoramic, multi-source evidence view: safety, efficacy, and PK/PD are assessed jointly in one ecosystem, eliminating the contradictions that arise when each domain is analysed in isolation. (3) From experience-driven judgment to model-driven quantitative inference: decisions are expressed as probabilistic statements — 'P(Go) = 72%, 95% credible interval [58%, 84%]' — rather than binary intuition ('I think this dose is safe'). These three shifts serve a single goal: converting data into defensible decisions: faster, more transparently, and more consistently.

OncoTrialMind is a production-grade implementation of the DDDM framework. It ingests CDISC-standard ADaM datasets (ADSL, ADAE, ADRS, ADTTE, ADTR) and user-uploaded PK exposure or competitive benchmark data through a unified reactive data layer, and organises over 40 analysis modules into five clinical function domains: Safety Analysis, Efficacy Analysis, PK/PD & Exposure-Response, Dose Optimization, and Decision Support. The platform deploys 35+ reactive server modules, supports four predefined user roles (Statistician, Medical Director, Project Manager, Auditor/Regulator), and provides complete audit-trail persistence to PostgreSQL. All Bayesian computation uses analytic conjugate forms or custom Metropolis-Hastings sampling in base R, with zero external MCMC compiler dependencies. This paper covers the system architecture, five-domain module organisation, underlying statistical and machine-learning methodology, and a case study illustrating the end-to-end decision workflow.

The remainder of this paper is organised as follows. Section 2 presents the system architecture (four-layer separation of concerns and four dashboard design principles). Section 3 provides an overview of the five-domain module map. Sections 4-8 detail each clinical function domain in turn: Safety Analysis (Section 4: knowledge graph, Landmark-TVC signal classification, CUSUM-Bayes monitoring, and AE clustering with risk stratification), Efficacy Analysis (Section 5), PK/PD & Exposure-Response (Section 6), Dose Optimization (Section 7), and Decision Support (Section 8: Go/No-Go scorecard, competitive benchmark & MAP prior, AI-augmented decision meeting report). Section 9 presents an end-to-end case study. Section 10 discusses statistical method validation, the GAMP 5 validation framework, and the decision audit trail. Section 11 concludes.

SYSTEM ARCHITECTURE

2.1 FOUR-LAYER SEPARATION OF CONCERNS

The architecture's core principle is four-layer separation of concerns.

The UI Layer (Shiny + bslib + Bootstrap 5, 10+ files in R/ui/) is responsible solely for presentation. Each module returns nav_panel() or nav_menu() as a pure function, enabling each UI domain to be independently developed, tested, and revised.

The Server Layer (Shiny reactive + promises::future_promise, 35+ files in R/server/) manages reactive computation and data-flow orchestration. All modules share a single cdisc_data() reactive data source via server_main.R's source(..., local = environment()) pattern.

The Statistical Layer (40+ pure R functions in R/*.R) implements all analytical algorithms with zero dependency on the Shiny environment. Every method can be independently tested and validated outside the application.

The Data Layer (PostgreSQL + DBI, R/db_connection.R) provides multi-user concurrent access, project-level data isolation, and an immutable audit trail.

A key design decision: adding a new module requires only declaring a dependency on cdisc_data, touching zero existing code.

Table 2-1. System Four-Layer Architecture

Layer	Technology Stack	File Location	Responsibility
UI Layer	Shiny + bslib + Bootstrap	R/ui/ui_*.R (10+ files)	Page layout, interactive controls — each module

	5		returns nav_panel() or nav_menu()
Server Layer	Shiny reactive + promises::future_promise	R/server/server_*.R (35+ files)	Reactive computation, data flow management — all modules share the same cdisc_data() source
Statistical Layer	R statistical functions	R/*.R (40+ analysis modules)	Pure functions — receive data, execute analysis, return results. Zero dependency on Shiny environment
Data Layer	PostgreSQL + DBI	R/db_connection.R	Data persistence, multi-user concurrent access, audit log storage

2.2 DASHBOARD DESIGN PRINCIPLES

All dashboards in the system follow four consistent design principles, grounded in human factors engineering and clinical decision-making practice.

(1) Progressive Disclosure. Information is organised in three tiers: L1 Overview Cards (1-3 key metrics with green/yellow/red colour coding for 5-second situational assessment); L2 Detail Panels (full charts, tables, and interpretive text); L3 Raw Data Drill-down (patient-level listings for audit and outlier review).

(2) Colour-Coded Signal Strength Tiering. All 40+ modules use a consistent three-colour scheme: Green (P(Go) >= 80%, proceed), Yellow (30-80%, intensified monitoring), Red (< 30%, immediate attention), supplemented with icons and positional encoding for colourblind accessibility.

(3) Cross-Module Reactive Linkage. Population filter conditions, dose selection (MTD/RP2D), and CDISC data uploads automatically propagate to all downstream modules through Shiny's reactive framework. A data selection made in one place takes effect across all views simultaneously.

(4) Role-Adaptive Decision Views. Four predefined roles share the same analysis engine: Statisticians access full parameter controls and diagnostic outputs; Medical Directors and Clinical Pharmacologists see decision-centric overviews with technical detail collapsed behind an Advanced panel; Project Managers focus on cross-project KPIs and milestone tracking; Auditors and Regulators can review decision traceability chains, audit logs, and data version snapshots.

MODULE OVERVIEW

The platform organises over 40 integrated analysis modules into five clinical function domains, each corresponding to a natural stage in the trial decision workflow. This domain-oriented organisation reflects how cross-functional development teams work in practice: safety reviewers focus on the Safety domain, pharmacometricians on PK/PD and E-R, clinical pharmacologists and statisticians on Dose Optimization, and governance committees on Decision Support. Domain outputs are not isolated; they feed into downstream domains through the reactive data layer, creating a cumulative evidence chain from raw CDISC data to the final Go/No-Go recommendation.

Table 3-1. Five Clinical Function Domains and Key Modules

Domain	Clinical Purpose	Key Modules
Safety Analysis	Detect, characterise, and monitor adverse events	AE Summary, DLT, EAIR, OCMQ, Knowledge Graph, Landmark Safety, TVC Cox, K-means AE Clustering, Logistic Regression Risk Scoring, CUSUM Monitoring

Efficacy Analysis	Assess tumour response and survival outcomes	ORR/DCR/DOR, Waterfall/Swimmer/Spider Plots, KM Survival (PFS/OS), Cox Regression, Subgroup Forest Plot
PK/PD & E-R	Model exposure-response relationships for dose justification	E-R Logistic/Emax/Loess, Bayesian E-R, Neural Network E-R, PK Quartile Survival Analysis, Treatment Window Decision
Dose Optimization	Identify MTD, OBD, and RP2D through adaptive designs	3+3, BOIN, CRM, mTPI, BOIN12, Combo BOIN, Simon Two-Stage, EffTox, BTTE, Seamless II/III, RP2D MDT
Decision Support	Synthesise cross-domain evidence into regulatory-grade decisions	Go/No-Go Scorecard (5-dim SMAA), BRAT B-R, AI Signal Engine (8 domains), Competitive Benchmark + MAP Prior, Decision Audit Trail, TPP/CDP Lifecycle

SAFETY ANALYSIS

The Safety domain addresses the most urgent question in early-phase oncology: is this molecule safe enough to continue development? Rather than walking through every standard AE table, we focus on three capabilities that elevate safety review from passive documentation to prospective intelligence: the AE co-occurrence knowledge graph, revealing hidden association structures among adverse events; the Landmark-TVC signal classifier, causally linking AE occurrence to survival outcomes; and the ML-driven CUSUM-Bayes coupled continuous safety monitoring system with AE clustering and patient risk stratification. Foundational descriptive analyses — incidence by PT/SOC, DLT rates per dose with exact binomial confidence intervals, exposure-adjusted incidence rates, and OCMQ grouping — are fully implemented and linked reactively to population filter, dose cohort, and biomarker subgroup selections.

4.1 AE KNOWLEDGE GRAPH: FROM ISOLATED EVENTS TO NETWORK TOPOLOGY

Traditional AE review treats each preferred term as an isolated count. In clinical practice, adverse events manifest as syndrome clusters — rash, colitis, hepatitis, and thyroiditis appearing together can signal immune-related adverse events (irAEs) characteristic of checkpoint inhibition; the triad of anaemia, neutropenia, and thrombocytopenia points to myelosuppression. R/knowledge_graph.R surfaces these latent patterns by constructing an interactive visNetwork PT-PT co-occurrence graph (Figure 1). Nodes are AE preferred terms sized by patient incidence and coloured by MedDRA SOC; edges connect PTs that co-occur in at least a user-specified minimum number of patients (default 2), weighted by co-occurrence count and annotated with Jaccard similarity. Node tooltips show affected patient count, Grade >= 3 events, SAEs, and drug-related events. Three layout algorithms (forceAtlas2Based, repulsion, barnesHut) reveal community structure; double-clicking any node isolates its 1-hop subgraph for focused review.

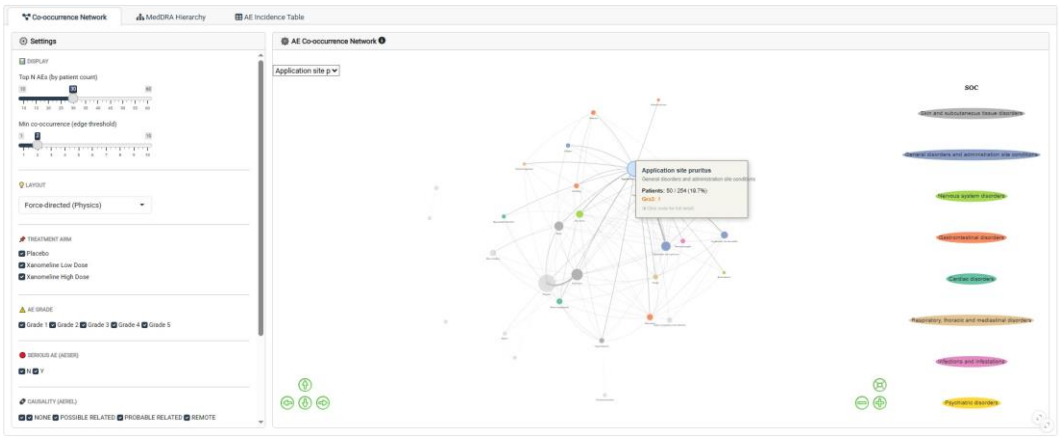

Figure 1. AE Knowledge Graph — interactive visNetwork PT-PT co-occurrence network.

Clicking a PT opens a detail panel showing CTCAE grade distribution (G1-G5), SAE rate, drug-related rate, and per-arm breakdown; clicking an edge shows co-occurring patient count, Jaccard coefficient, and per-arm co-occurrence distribution.

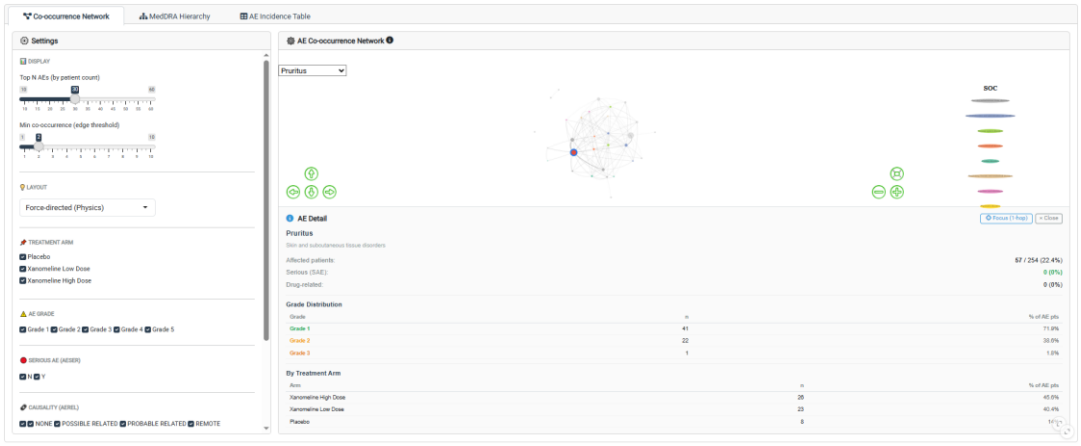

Figure 2. PT node click detail panel: CTCAE grade distribution (G1-G5), SAE rate, drug-related rate, and per-arm breakdown.

The network perspective offers three decision advantages. First, early syndrome recognition: a spontaneously emerging immune-related AE cluster provides objective evidence for establishing irAE management guidelines before individual PT incidence crosses conventional thresholds. Second, leading indicators of dose-limiting toxicity: certain AE co-occurrence patterns — the 'fatigue + decreased appetite + weight loss' triad clustering at a given dose level — may signal proximity to the maximum tolerated dose even when no single PT has reached the 33% DLT boundary. Third, distinguishing on-target from off-target toxicity: comparing the trial drug's network topology against known class-effect patterns (for example, EGFR inhibitors produce a characteristic skin-paronychia-diarrhoea cluster): deviation from this expected topology warrants investigation. The knowledge graph thus transforms safety review from retrospective tabulation ('what happened?') to prospective pattern recognition ('what is emerging?'), feeding directly into the Decision Support domain's benefit-risk synthesis.

4.2 AE-TO-SURVIVAL SIGNAL CLASSIFICATION: LANDMARK + TIME-VARYING COX

A fundamental question in oncology safety review is whether experiencing a treatment-emergent AE predicts subsequent deterioration in progression-free survival. Answering this question is complicated by immortal time bias: patients must survive long enough to experience the AE, so a naive comparison of 'AE occurred' vs. 'did not occur' systematically favours the AE group, because the survival time before AE onset is incorrectly attributed to the AE itself. R/ae_tvc_analysis.R addresses this bias through two

complementary methods. The Landmark analysis (Figure 3) sets a user-specified landmark time point, excludes patients who experienced PFS events or were censored before the landmark, stratifies the remaining patients by whether the target AE occurred before the landmark, and resets the time origin to the landmark day: post-landmark survival comparisons are thus freed from immortal time contamination. The resulting Kaplan-Meier plot displays AE+ vs AE- curves, Cox HR, 95% CI, p-value, and an at-risk table, enabling complete interpretation without switching views.

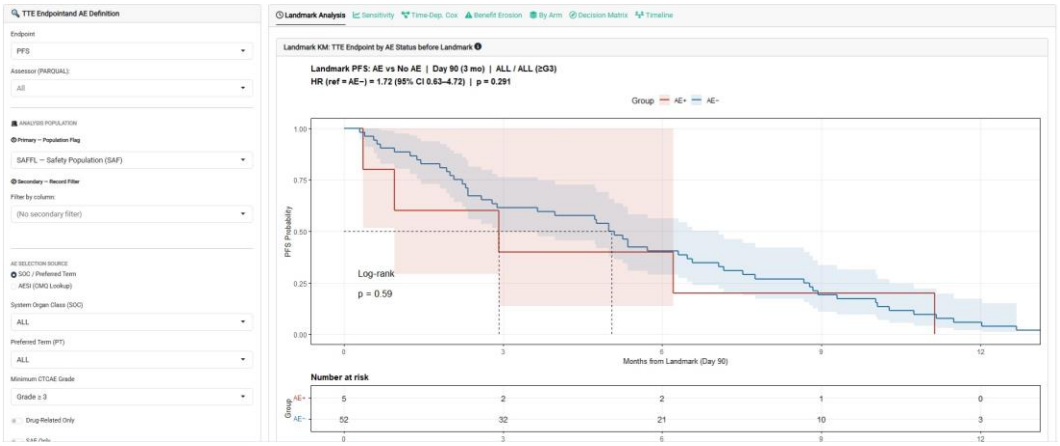

Figure 3. Landmark Kaplan-Meier PFS curves stratified by AE occurrence before the landmark time point, with Cox HR, 95% CI, and at-risk table.

In Figure 3's Landmark plot, the blue curve represents patients who experienced the target AE before the landmark (AE+ group), the red curve those who did not (AE- group). The degree of separation between the two curves directly reflects the strength of association between the AE and subsequent PFS. The shaded regions are 95% confidence bands, whose width expands as the at-risk count declines, visually conveying estimation uncertainty in the tail. The upper-left annotation reports three key statistics: the Cox HR quantifies the hazard ratio of AE+ relative to AE- (HR > 1 indicates higher progression risk in the AE+ group), the 95% CI provides estimation precision, and the log-rank p-value gives statistical significance. The at-risk table at the bottom lists, by time point, the number of patients still at risk in each group, used to judge the reliability of tail estimates. When at-risk counts drop to single digits, tail fluctuations should be interpreted cautiously. If the curves show persistently lower PFS in the AE+ group with a significantly elevated HR (e.g., HR = 2.1, p = 0.003), this suggests the AE may be an early signal of drug toxicity; conversely, if the AE+ group shows better PFS than the AE- group (HR < 1), one should consider whether the AE may serve as a pharmacodynamic biomarker: precisely the statistical validation of the knowledge graph's on-target effect identification described in Section 4.1.

The time-varying covariate (TVC) Cox model complements the Landmark approach: AE status is treated as a time-dependent exposure Z(t), where patients contribute person-time to the "AE-free" risk set until the AE occurs, then switch to the "AE-occurred" risk set. Unlike Landmark, which reduces AE timing to a binary variable before a single cut-point, TVC preserves the exact AE onset time and captures events occurring after the landmark. The classify_ae_outcome_signal() function integrates both models into an eight-category decision framework (Table 4-1), with results displayed as styled cards in the Decision Matrix sub-tab, each card carrying a category-specific clinical interpretation and actionable recommendations (Figure 4).

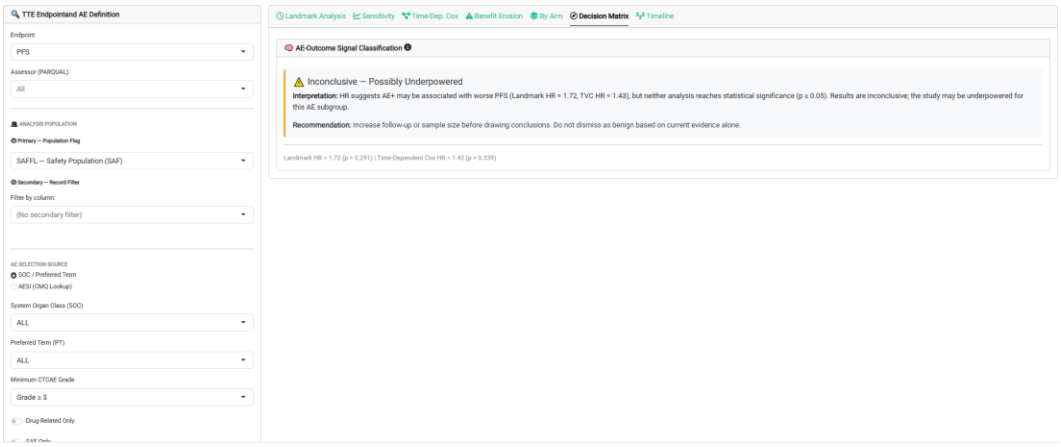

Figure 4. Decision Matrix result card: classification outcome with colour-coded severity border, clinical interpretation, and recommended action.

By cross-referencing Landmark and TVC significance status, HR direction, and between-model consistency, the classification framework transforms two sets of statistical output into structured clinical decision guidance: distinguishing true drug-specific toxicity from disease-driven events, consistent on-target effects from exploratory single-method signals, and acute from cumulative toxicity patterns. Toxicity-Driven, Delayed Cumulative, and Early Predictor signals (Table 4-1 rows 3, 7, 8) reduce the Go/No-Go efficacy score; On-Target Effect signals (rows 4, 6) increase it, wiring safety-to-survival evidence directly into the governance decision calculus.

Landmark ($p < 0.05$)	TVC ($p < 0.05$)	HR > 1	HR < 1	Classification	Clinical Interpretation
Significant	Significant	Yes	—	Toxicity-Driven Outcome Signal	AE consistently associated with worse PFS in both models. Investigate dose interruption/reduction patterns; consider dose optimization.
Significant	Significant	—	Yes	On-Target Effect (Consistent)	AE consistently associated with improved PFS. Likely surrogate biomarker of drug activity; manage tolerably, do not over-suppress.
Significant	Significant	Yes (TVC)	Yes (LM)	Discordant Signal	Contradictory HR directions between models. Possible time-varying effect or informative censoring; conduct subgroup analysis.
Not significant	Significant	Yes	—	Delayed Cumulative Signal	Early AE does not predict, but time-varying exposure significantly worsens PFS. Suggests chronic cumulative toxicity; extend

					safety follow-up.
Significant	Not significant	Yes	—	Early Predictor (Limited Duration)	AE before landmark predicts worse PFS, but no ongoing cumulative effect. Use landmark AE status as a stratification factor.
—	—	—	Yes	Possible On-Target (Single Method)	Only one method shows protective association. Exploratory signal; requires larger sample to confirm.
Not significant	Not significant	≈1	≈1	Benign AE	AE not associated with PFS/OS (both models HR ≈ 1, $p \geq 0.05$). True drug-specific toxicity; standard safety management.
Not significant	Not significant	≠1	≠1	Inconclusive (Underpowered)	HR suggests directional trend but neither method reaches significance. Sample size may be inadequate; extend follow-up before concluding.

4.3 ML SAFETY MONITORING: ADAPTIVE CUSUM-BAYES COUPLING

Traditional DLT monitoring uses binary logic (a patient either has a DLT or does not), creating a blind spot: patients who accumulate multiple Grade 2-3 events, experience dose modifications, or develop SAEs short of the formal DLT threshold remain invisible until the boundary is crossed. The ML Safety Strategy module, located under Intelligent Decision Support > ML Safety Strategy, closes this gap with continuous patient-level risk surveillance (Figure 5). Seven AE-derived features, weighted by pre-specified severity coefficients, are fed into an XGBoost classifier to produce a per-patient risk score with SHAP waterfall-plot interpretability; when data are insufficient, the model falls back to a linear clinical prior. The key innovation is the adaptive coupling mechanism in run_safety_cusum(): the CUSUM reference rate p0 is dynamically set from the Bayesian DLT posterior median, making the sequential monitor more sensitive at high-risk dose levels and more permissive at safer doses: a feedback loop between model-based posterior and model-free sequential monitoring. An integrated utility score with adaptive efficacy/toxicity weights produces a single five-tier recommendation (Strong Go to No-Go), displayed alongside the triggering evidence in a human-in-the-loop design.

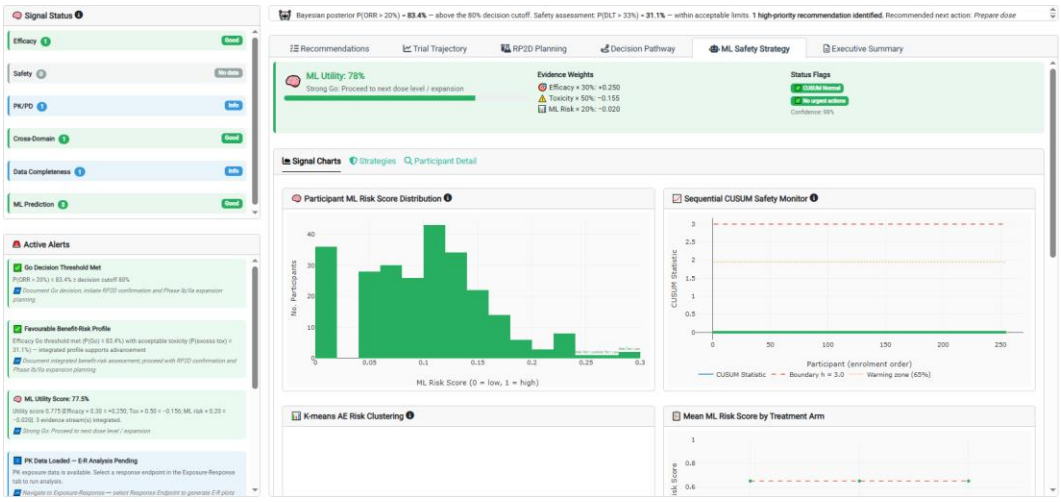

Figure 5. ML Safety Strategy interface: utility score banner with Bayesian posteriors and dynamic evidence weights; participant ML risk score distribution (left); Sequential CUSUM Safety Monitor with adaptive decision boundary coupled to DLT posterior (right).

Continuous ML surveillance offers three decision advantages. First, it catches the 'rising risk' patient — someone with multiple moderate events who has not crossed the DLT boundary but is trending toward it — enabling pre-emptive dose modification rather than waiting for passive DLT reporting. Second, the CUSUM-Bayes coupling relieves medical monitors of the burden of manually adjusting monitoring thresholds when moving between dose levels: at high-risk doses, alarms trigger earlier; at low-risk doses, false alarms are suppressed; the adjustment is data-driven, not intuitive. Third, when the Bayesian DLT posterior and ML risk stratification independently and simultaneously signal safety concern — two orthogonal evidence streams converging — this is the strongest warning short of an actual SAE cluster, triggering immediate benefit-risk reassessment. The module thus transforms safety monitoring from episodic DLT counting to continuous risk anticipation, directly feeding the Decision Support domain's benefit-risk synthesis.

EFFICACY ANALYSIS

The Efficacy domain provides ORR, DCR, and DOR from ADRS data; interactive Waterfall, Swimmer, and Spider plots via plotly; Kaplan-Meier PFS/OS estimation with Cox regression; and a Subgroup Forest Plot for covariate-stratified comparisons. These are well-established methods. What distinguishes the platform is a sequential analytical logic that mirrors the clinical reasoning chain: first, assess the quality of tumour response — does it last? (ORR + DOR joint analysis); second, confirm whether response translates into survival benefit — does it matter clinically? (ORR + PFS joint analysis). Each step builds on the previous, and only when both are satisfied does the efficacy evidence support a high-confidence Go decision.

5.1 ORR + DOR JOINT ANALYSIS: THE QUALITY OF RESPONSE

ORR measures whether the drug can shrink tumours; DOR measures whether that shrinkage is durable. The platform captures this through two complementary metrics. The Durable Response Rate (DRR) — defined as the proportion of evaluable patients achieving CR or PR with DOR exceeding a pre-specified threshold (e.g., 6 months) — directly supports FDA accelerated approval evidence requirements by distinguishing transient radiographic change from clinically meaningful tumour control. The Bayesian ORR x mDOR joint posterior goes further, modelling ORR and median DOR as a bivariate random variable (π, μ). ORR follows the Beta-Binomial conjugate model; mDOR follows an Exponential-Gamma model (log(2)/λ, where λ ~ Gamma(a0 + r, b0 + Σ t_i), conserving conjugacy at the cost of the constant-hazard assumption). Monte Carlo draws from the joint posterior are plotted in two dimensions, overlaid with the TPP dual-target rectangle (ORR >= θ_orr AND mDOR >= θ_dor) (Figure 6). The proportion of draws falling in this region is the joint posterior probability.

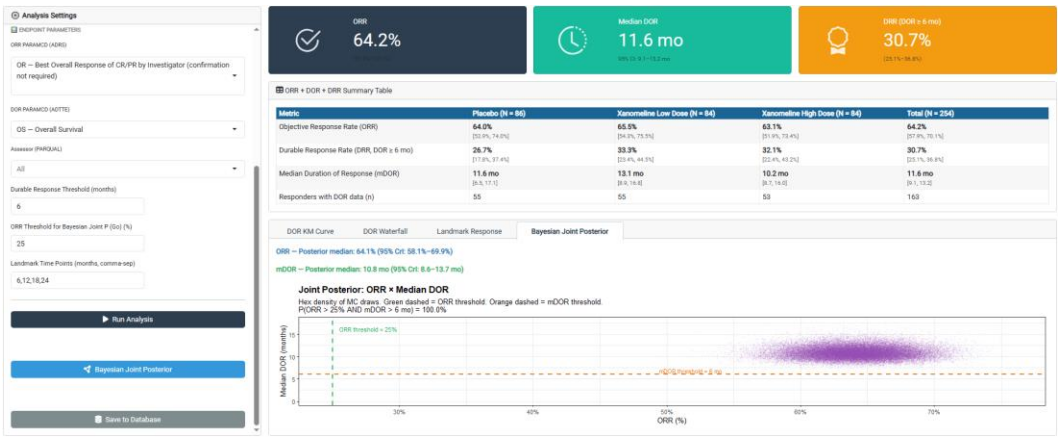

Figure 6. ORR x mDOR joint posterior distribution: ORR summary cards and arm-wise comparison table (top); bivariate Monte Carlo posterior with TPP dual-target thresholds and joint Go probability (bottom).

Each purple point in the joint posterior plot represents one Monte Carlo draw from the bivariate posterior — a plausible (ORR, mDOR) pair given the observed data — with darker regions indicating higher posterior density. The green dashed vertical line marks the TPP ORR threshold (θ_orr); the orange dashed horizontal line marks the TPP mDOR threshold (θ_dor, e.g., 6 months). Together they partition the plane into four quadrants, the upper-right being the dual-target pass zone. The annotation in the upper-right corner reports the proportion of draws falling in that zone — P(ORR > θ_orr AND mDOR > θ_dor | Data) — the joint Go probability. In the example shown, 100% of draws lie in the dual-target zone, representing the strongest possible efficacy signal: regardless of posterior uncertainty in ORR or mDOR individually, every plausible parameter combination simultaneously satisfies both TPP targets. This is the ideal profile for initiating Phase III or accelerated approval submission.

The decision value of the joint analysis is its ability to discriminate three clinically distinct patterns that identical single-endpoint ORR cannot separate. (1) Dual-target pass (high joint probability, e.g., > 80%): both ORR and mDOR satisfy TPP targets — Figure 6 is a textbook example. (2) One target passes, the other is borderline (joint probability 40-60%): the drug induces responses but they are not durable — the classic 'high-ORR short-DOR' pattern that explains many Phase II positive / Phase III negative failures. This triggers extended follow-up or the ORR + PFS joint analysis (§5.2) to assess whether responses nonetheless translate into survival benefit. (3) Dual-target fail (joint probability < 30%): strong evidence against continuation at the current dose. By encoding both dimensions in a single probability, the joint posterior prevents the governance error of approving on ORR alone without establishing whether responses persist. The analysis is implemented in R/orr_tte_joint.R, with the Shiny interface in R/server/server_orr_tte.R.

5.2 ORR + PFS JOINT ANALYSIS: BRIDGING RESPONSE TO SURVIVAL

If ORR + DOR analysis answers 'is the response meaningful?', then ORR + PFS analysis answers 'does the response translate into survival benefit?' This is the most clinically consequential question in early-phase efficacy assessment, directly addressing the evidentiary standard for FDA accelerated approval: can the ORR signal predict PFS benefit in a confirmatory trial?

The analysis proceeds through a two-step evidence chain. Step 1 — durability: are responses sustained (via DRR or ORR x DOR joint posterior, §5.1)? Step 2 — translation: do responders experience significantly longer PFS than non-responders? Step 2 is assessed by Landmark responder PFS analysis: at a fixed landmark time, patients still at risk are stratified by whether they achieved CR/PR before the landmark, and post-landmark PFS is compared via Kaplan-Meier and Cox regression. This framework uses the same immortal-time-bias correction as the AE-to-PFS analysis (§4.2), but here evaluates whether tumour response predicts subsequent disease control.

The Bayesian ORR x PFS joint posterior extends the bivariate framework of §5.1 to the response-survival

pair: ORR ~ Beta-Binomial, mPFS ~ Gamma-Exponential (λ ~ Gamma(a0 + d, b0 + T_total), where d is the total number of PFS events and T_total is total follow-up person-months), with Monte Carlo integration computing P(ORR > θ_orr AND mPFS > θ_pfs | Data) (Figure 7).

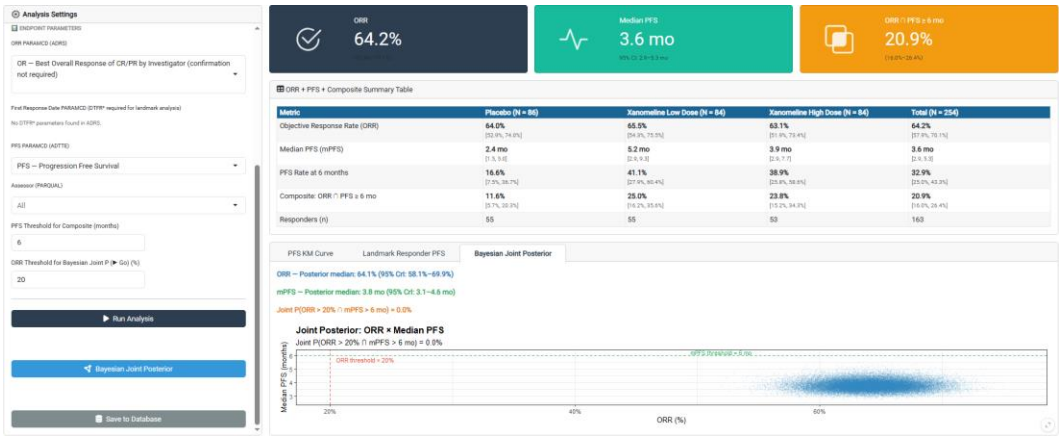

Figure 7. ORR x PFS joint posterior distribution: ORR and PFS summary cards with composite endpoint (top); arm-wise comparison table (middle); bivariate Monte Carlo posterior with TPP dual-target thresholds and joint Go probability (bottom).

Each blue point in the joint posterior plot represents one Monte Carlo draw from the bivariate posterior — a plausible (ORR, mPFS) pair given the observed data. The red dashed vertical line marks the TPP ORR threshold; the green dashed horizontal line marks the TPP mPFS threshold (e.g., 6 months). The upper-right quadrant is the dual-target pass zone, and the annotation reports P(ORR > θ_orr AND mPFS > θ_pfs | Data). In the example shown, the scatter cloud is concentrated in the lower-right region: ORR is high (~64%, well above the 20% threshold), but mPFS hovers around 3-4 months, far below the 6-month threshold. The joint probability is 0.0% — not low, but zero. This is the classic 'high-ORR short-PFS' trap: tumour shrinkage is real, but it does not translate into sustained disease control. If decision-makers looked only at ORR, they would see a 64% response rate and conclude the drug works; the joint posterior reveals what ORR alone conceals — that not a single Monte Carlo draw satisfies both TPP targets simultaneously.

The two-step chain provides structured decision logic: both steps passing provides the strongest possible early-phase efficacy evidence for a Go decision; high ORR with poor DOR and no responder PFS advantage — as in Figure 7 — strongly suggests ORR is a noise signal, the root cause of many Phase III failures. The joint posterior translates this into a single auditable probability, directly supporting the governance question: "do we have evidence that tumour shrinkage translates into longer progression-free survival?" The analysis is implemented in R/orr_tte_joint.R (compute_orr_pfs_joint_posterior), with Shiny interfaces in R/server/server_orr_tte.R and R/server/server_bayes.R.

PK/PD AND EXPOSURE-RESPONSE ANALYSIS

The E-R domain links drug exposure (AUC, Cmax, Ctrough) to both efficacy and safety endpoints via R/exposure_response_ml.R (2,979 lines, 20+ functions). The modelling suite spans Logistic regression, Sigmoid Emax, non-parametric Loess, Bayesian E-R with full posterior quantification, and neural-network-based E-R with XGBoost multi-feature extensions, with automated model selection by cross-validated AUC. The entire modelling effort converges on a single decision output: the therapeutic window.

6.1 TREATMENT WINDOW: FROM E-R ANALYSIS TO DOSE RECOMMENDATION

The platform fits separate efficacy-E-R and toxicity-E-R curves using the modelling suite, then overlays them on the same exposure axis. Two horizontal reference lines — the TPP minimum acceptable ORR and the maximum tolerable Grade 3+ AE rate — partition the plane; the green shaded region between them is the therapeutic window, where efficacy is sufficient and toxicity is acceptable. A dose-exposure summary table (mean, geometric mean, CV%, quartiles of AUC by dose group) maps the recommended exposure range back to candidate dose levels. The joint overlay produces four decision patterns. (1)

Wide therapeutic window: positive efficacy E-R with flat toxicity E-R — the ideal case, allowing dose selection that maximises efficacy with no toxicity penalty. (2) Narrow therapeutic window: both curves positive — requiring precise exposure control and formal benefit-risk trade-off (§8). (3) Toxicity-dominated: flat efficacy but positive toxicity E-R — the minimum effective dose has been exceeded; reduce dose. (4) No E-R signal: neither curve informative within the explored range — expand the dose range. When the efficacy and toxicity boundaries cross — efficacy requiring higher exposure than safety permits — the window closes entirely, representing a fundamental liability of the molecule.

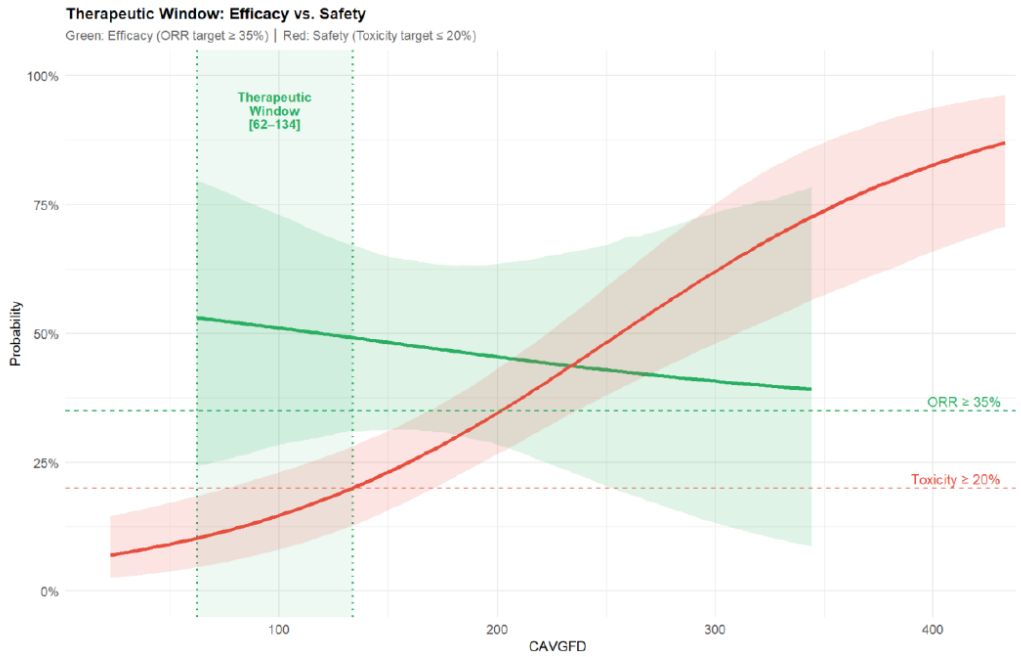

Figure 8. Joint E-R curves and therapeutic window: green efficacy E-R curve and red toxicity E-R curve overlaid on the same exposure axis, with two horizontal reference lines (TPP minimum acceptable ORR, maximum tolerable G3+ AE rate) delineating the green-shaded therapeutic window region — where efficacy is sufficient and toxicity is acceptable.

What makes the therapeutic window a decision tool rather than a curve-fitting exercise is Bayesian posterior quantification. Classical maximum-likelihood E-R produces point estimates — a single efficacy curve and a single toxicity curve — but cannot answer whether an apparent gap between them reflects a genuine therapeutic window or sampling noise. Bayesian E-R replaces deterministic fitted lines with posterior predictive curves carrying 95% credible bands. These bands widen at exposure tails, making data sparsity directly visible; when the credible band at the proposed RP2D exposure lies entirely above the efficacy threshold and entirely below the safety ceiling, the window is probabilistically confirmed. The Bayesian E-R implementation uses rstanarm for Logistic models and brms for Emax models, with a pure-R Metropolis-Hastings fallback that ensures the module deploys without external compiler dependencies. At the decision exposure, the posterior probability $P(ORR > \theta_0 \mid \text{exposure} = x^*)$ maps to a three-tier Go/Gray-Zone/No-Go classification, establishing a consistent probability vocabulary with the platform-wide Bayesian decision framework.

The E-R domain's ultimate output is not a set of coefficients and p-values but a single, quantified dose recommendation: 'Recommended RP2D = X mg, with predicted ORR = Y% and Grade 3+ AE rate = Z%. Rationale: efficacy has reached 90% of the Emax plateau while toxicity remains within the acceptable ceiling; doubling the dose would increase ORR by only 3% but raise Grade 3+ AE rate by 15%.' This is the most important E-R output for governance discussions. It translates PK modelling into a binary, clinically interpretable question — does a therapeutic window exist? — and answers it with probabilistic rigour. The E-R results feed directly into the Decision Support domain's benefit-risk synthesis (§8), where the therapeutic window determination serves as a core evidence input to the BRAT multi-criteria

framework. The analysis is implemented in R/exposure_response_ml.R (fit_er_bayes_logistic, fit_er_bayes_emax), with Shiny interfaces in R/server/server_er.R and R/server/server_exposure.R.

DOSE OPTIMIZATION

The Dose Optimization domain covers the complete dose-finding journey from first-in-human to registrational readiness.

Phase I dose escalation provides four methods with full operating characteristic simulation: 3+3 (historical reference), BOIN (Liu & Yuan 2015, closed-form boundaries), mTPI (Ji et al. 2010, unit probability mass), and Bayesian CRM with custom Metropolis-Hastings sampling (4 chains x 20,000 iterations, EWOC safety constraint). Combination dose escalation (R/combo_dose_escalation.R) extends BOIN to two-dimensional dose-toxicity matrices.

For Phase II, Simon Two-Stage design (R/simon_two_stage.R) uses exhaustive search with three-layer pruning for sub-5-second execution, returning Optimal and Minimax solutions with Koyama & Chen (2008) adjusted p-values; Joint Efficacy-Toxicity Phase II (EffTox, Thall, Simon & Estey 1995; R/efftox_ph2.R) and Bayesian TTE Phase II (Exponential-Gamma conjugate) complete the single-stage Phase II toolkit. Seamless Phase II/III design (R/seamless_ph23.R) unifies exploration and confirmation into a single continuous trial, with Winner's Curse bias correction and Fisher combination test multiplicity control.

The following three sections focus on three key innovations spanning the Phase I-II development continuum: BOIN12 (optimal biological dose finding), Simon Two-Stage (Phase II futility monitoring), and Eff/Tox (joint efficacy-toxicity sequential decision-making).

7.1 BOIN12: UTILITY-BASED OPTIMAL BIOLOGICAL DOSE FINDING

FDA's Project Optimus initiative has fundamentally shifted the goal of dose finding: no longer the highest tolerated dose, but the dose that optimally balances efficacy and toxicity. BOIN12 extends the BOIN framework to jointly model efficacy and toxicity, assigning each patient outcome to one of four ordinal categories (no DLT + no efficacy; no DLT + efficacy; DLT + no efficacy; DLT + efficacy) and computing a utility score that quantifies the desirability of each dose level. The dose maximising expected utility is the Optimal Biological Dose (OBD) (Figure 9). By providing a transparent, pre-specified utility function, BOIN12 enforces a critical decision discipline: the utility weights, target DLT rate, and utility threshold must be declared before the trial begins, precluding retrospective dose selection based on post-hoc preference.

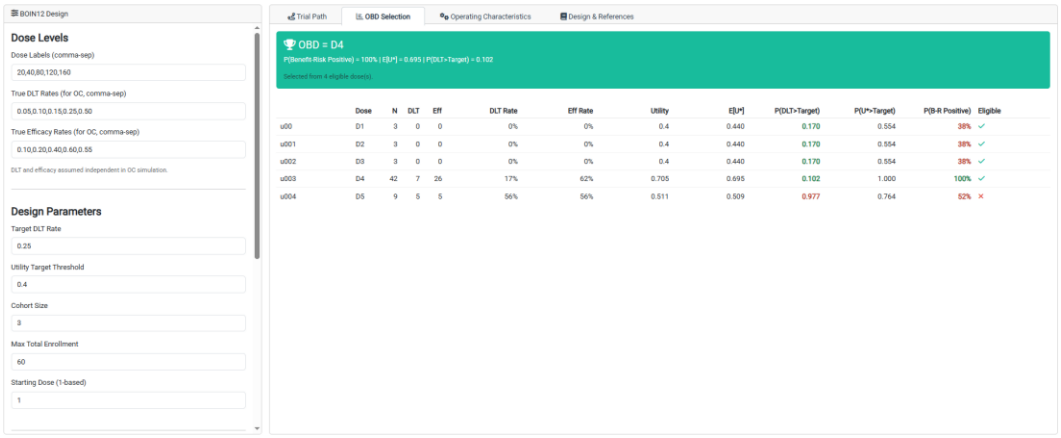

Figure 9. BOIN12 OBD Selection: per-dose decision table with enrolled patients, observed DLT/efficacy event counts, DLT rate, efficacy rate, utility score, and eligibility flag; green check = OBD, red cross = eliminated.

The OBD Selection table (Figure 9) lets the decision-maker see not just which dose won, but why every other dose lost. In the example shown, five dose levels (20-160 mg) are evaluated under a target DLT rate of 25% and a utility threshold of 0.40. BOIN12 adaptively concentrated enrolment on the most promising dose: 42 of 60 patients were assigned to Dose 4 (120 mg), which achieved a 62% efficacy rate with a 17% DLT rate — comfortably below the 25% target — yielding a utility of 0.705 and a posterior P(Benefit-Risk Positive) of 100%. Dose 5 (160 mg) was eliminated: despite 56% efficacy, its DLT rate of 56% far exceeded the safety boundary, and its benefit-risk probability fell to 52%. Doses 1-3, while safe, accumulated too few patients to compete on utility. This transparent, data-driven ranking transforms dose selection from an implicit judgment — 'Dose 4 looks about right' — into an auditable decision: 'Dose 4 maximises expected utility, with 100% posterior probability of positive benefit-risk.' The module is implemented in R/boin12.R, with the Shiny interface in R/server/server_boin12.R.

7.2 SIMON TWO-STAGE: OPTIMAL FUTILITY MONITORING FOR PHASE II

Simon's Two-Stage design (Simon 1989) addresses the most consequential Phase II question: 'When should a futile trial be stopped early?' In oncology Phase II, most experimental drugs fail to achieve sufficient single-agent activity to warrant Phase III investment, yet under a single-stage design this determination can only be made after the full sample is enrolled. Simon's design provides a formal framework for early futility stopping — after Stage 1, if the observed number of responses falls at or below a pre-specified boundary r1, the trial stops for futility, protecting patients from continued exposure to an ineffective drug and conserving development resources. The design requires four parameters: alpha (Type I error), beta (Type II error), p0 (unacceptable response rate floor), and p1 (target response rate worth detecting). The platform implements both optimisation strategies via exhaustive search with three-layer pruning. The Optimal design minimises E(N | p = p0) — the expected sample size when the drug is inactive — recognising that most Phase II drugs fail and expected patient exposure to futile treatment should be minimised. The Minimax design minimises the total sample size n, providing a guaranteed upper bound on enrolment regardless of whether the trial stops at Stage 1 or proceeds to Stage 2. Both designs return n1, n2, r1, and r, accompanied by operating characteristic curves and Koyama & Chen (2008) adjusted p-values (Figures 10-11).

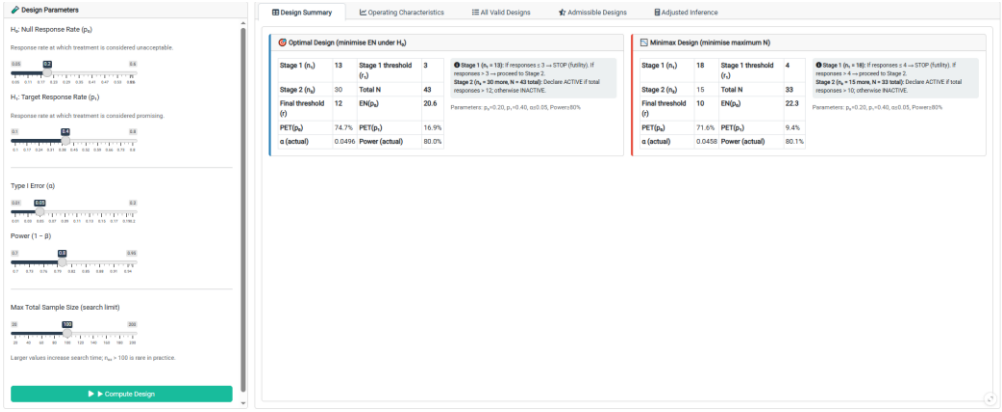

Figure 10. Simon Two-Stage Design: interactive design tool showing parameter inputs (p0, p1, alpha, power) on the left and the two computed design solutions on the right — Optimal Design (minimising expected sample size under the null) and Minimax Design (minimising maximum total sample size).

Figure 10 presents the Simon interactive design tool with parameter inputs on the left and two computed solutions on the right. The left panel collects the four core parameters: p0 = 0.20 (unacceptable response rate floor), p1 = 0.40 (target response rate), alpha = 0.05, and power = 0.80. The right panel displays the Optimal and Minimax designs side by side for direct comparison. In this example, the Optimal design requires n1 = 13 in Stage 1 with an early stopping boundary of r1 <= 3 (i.e., stop for futility if three or fewer responders), followed by n2 = 30 in Stage 2, yielding a total n = 43 with a final Go threshold of r > 12. The expected sample size under the null is E(N | p0) = 20.6 — if the drug is truly inactive, the trial stops after approximately 21 patients on average. The Minimax design shifts the allocation: n1 = 18, r1 <= 4, n2 = 15, total n = 33, with E(N | p0) = 22.3. The trade-off is clear: Optimal saves more patients on

average when the drug is ineffective (20.6 vs. 22.3), but Minimax guarantees a lower worst-case enrolment ceiling (33 vs. 43). Both designs achieve approximately 80% power. The choice between them is not purely statistical but depends on context. For most early-phase oncology trials, where the majority of experimental drugs fail, the higher probability of early termination under futility makes the Optimal design the right default. For ultra-rare indications — where recruiting 10 additional patients could add years to the trial — or when the drug has a high prior probability of success, the Minimax design is preferred because its lower absolute enrolment ceiling keeps the worst case contained. A useful shorthand: Optimal prioritises expected cost under futility; Minimax guarantees an absolute enrolment cap. When in doubt, go Optimal.

Figure 11. Simon Two-Stage operating characteristic curves: P(declare active) and expected sample size E(N) as functions of true response rate, with Optimal (solid blue) and Minimax (dashed red) overlaid for comparison.

Figure 11 shows the operating characteristic curves that validate the design's statistical performance across the full range of possible true response rates. The upper panel plots P(declare active) — the probability the trial concludes the drug is effective — as a function of the true response rate p. The curve rises from near alpha = 0.05 at p = p0 = 0.20 (the design controls Type I error) to approximately 0.80 at p = p1 = 0.40 (the design achieves nominal power). The two vertical dashed lines mark p0 and p1, creating an intuitive visual anchor: the curve should be low at the left marker and high at the right marker. The Optimal (solid blue) and Minimax (dashed red) curves are nearly indistinguishable, confirming that both designs provide equivalent statistical power. The lower panel plots expected sample size E(N) as a function of p, revealing where the two designs genuinely differ. At low true response rates (p < 0.25), the Optimal design's blue curve sits visibly below the Minimax red curve — Optimal stops earlier when the drug is ineffective. This OC visualisation transforms abstract alpha-beta guarantees into a concrete performance profile that clinical teams and regulators can evaluate before the first patient is enrolled: the upper panel confirms error control, the lower panel reveals the sample-size trade-off.

Simon Two-Stage's core governance value is pre-specification discipline. Every parameter — p0, p1, alpha, beta, and the resulting n1, n2, r1, r boundary values — must be committed before the first patient is enrolled. Once Stage 1 completes, the decision rule is deterministic: count responders, compare to r1. If responders <= r1, the trial stops — no discussion, no post-hoc interpretation. If responders > r1, the trial proceeds to Stage 2. This transparency eliminates the most common source of Phase II controversy: continuing a trial that the pre-specified criteria say should stop, because the observed ORR 'looks promising enough' or a subgroup analysis 'suggests benefit in a particular subgroup.' Such post-hoc reasoning, however well-intentioned, undermines the Type I error control that the design was built to provide. The module is implemented in R/simon_two_stage.R with exhaustive search and three-layer pruning for sub-5-second execution, and the Shiny interface in R/server/server_simon.R.

7.3 EFF/TOX SEQUENTIAL MONITORING: JOINT EFFICACY-TOXICITY DECISION FRAMEWORK

Simon's Two-Stage design addresses one critical Phase II question — futility — but it has a recognised blind spot: it evaluates efficacy alone, ignoring toxicity. A drug with ORR = 60% but Grade 3+ AE rate =

50% would pass a Simon design with flying colours, yet in clinical reality such a drug would never reach Phase III — its toxicity profile would be unacceptable in any benefit-risk assessment. The Eff/Tox sequential monitoring framework (Thall, Simon & Estey 1995) closes this gap by formally incorporating both efficacy probability pE and toxicity probability pT into a single decision architecture. The platform implements two independent Beta-Binomial conjugate posteriors with Jeffreys priors Beta(0.5, 0.5), evaluated at a series of pre-specified interim sample sizes (default N = 20 / 35 / 50). At each interim, the updated posteriors drive a four-region decision framework with strict priority ordering (Figures 12-14). Priority 1 — Excessive Toxicity Stop: if P(pT >= λ_T | Data) >= η_T (default 0.80), the trial stops regardless of efficacy. Priority 2 — Futility Stop: if P(pE <= λ_E | Data) >= η_F (default 0.80) and toxicity has not already stopped the trial, terminate for insufficient efficacy. Priority 3 — Success Declaration: if P(pE > λ_E) >= η_E (default 0.80) AND P(pT < λ_T) >= tox_go_cutoff (default 0.50), declare the drug promising. Priority 4 — Continue Enrollment: if none of the above conditions are met, the evidence is insufficient and enrollment proceeds to the next interim. Critically, this priority ordering is not a user-configurable option — it is enforced through the sequential if-else structure of stop_toxicity preceding stop_futility preceding declare_go in the efftox_ph2.R implementation.

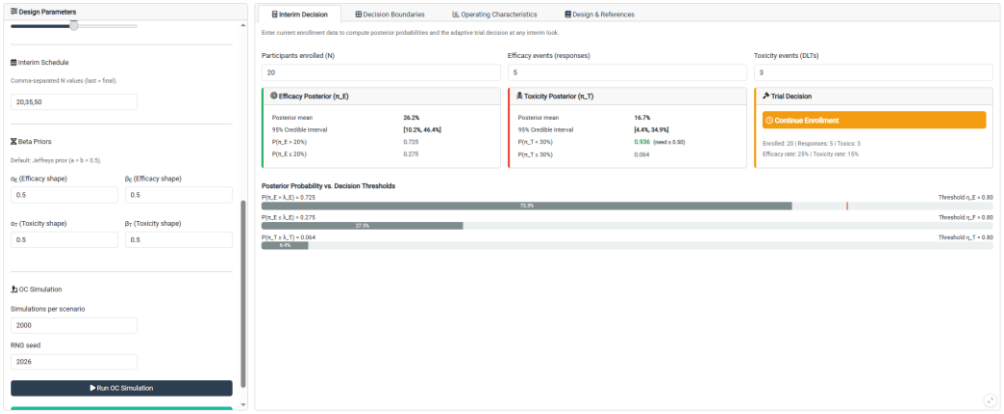

Figure 12. Eff/Tox real-time decision interface: interim schedule and prior configuration (left), posterior probability calculation for efficacy (green) and toxicity (red) at the current interim N = 20 with 5 responders and 3 DLTs, probability-versus-threshold gauge chart, and the trial decision output (Continue Enrollment).

Figure 12 shows the Eff/Tox interactive decision tool at the first interim analysis (N = 20). The left panel defines the design parameters: interim schedule at N = 20 / 35 / 50, Jeffreys priors Beta(0.5, 0.5) for both efficacy and toxicity, and OC simulation settings (2,000 replicates per scenario, seed 2026). The right panel presents the live posterior calculation after 20 patients with 5 efficacy events and 3 DLT events. The green efficacy posterior yields a mean pE of 26.2% with a 95% credible interval of [10.2%, 46.4%], and P(pE > 20%) = 0.725 — the drug has a 72.5% probability of exceeding the 20% efficacy floor. The red toxicity posterior yields a mean pT of 16.7% with P(pT < 30%) = 0.936 — toxicity is comfortably below the 30% ceiling with 93.6% confidence. The probability-versus-threshold gauge chart at bottom shows the efficacy probability (0.725) falling short of the 0.80 success threshold while the toxicity probability comfortably clears its boundary. The trial decision panel outputs 'Continue Enrollment' — efficacy is directionally positive but not yet decisive, safety is strong, so the system recommends proceeding to N = 35. This is Bayesian adaptive logic in action: not a binary Go/Stop at a fixed boundary, but a probabilistic evidence-strength assessment that acknowledges gradations of certainty.

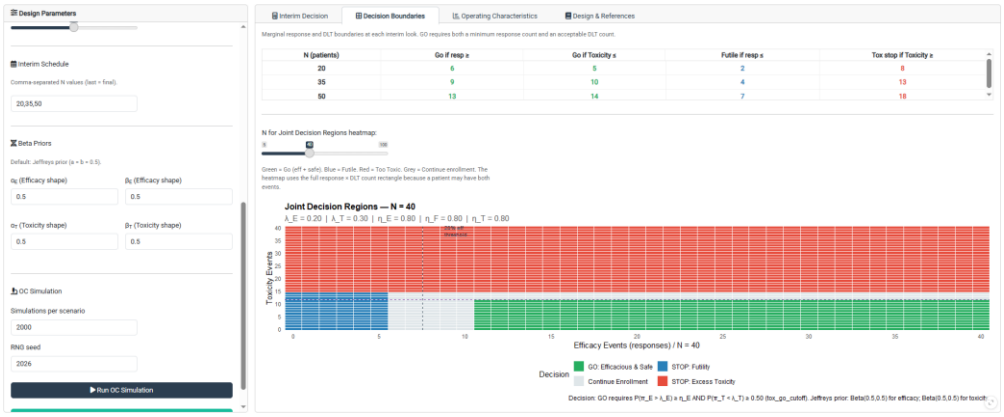

Figure 13. Eff/Tox decision boundary visualisation: per-interim threshold table (Go, Futility, Tox Stop boundaries at N = 20/35/50) and joint decision region heatmap in (n_eff, n_tox) space, colour-coded by decision outcome: green = GO, blue = futility stop, red = excess toxicity stop, grey = continue.

Figure 13 translates the probabilistic framework into concrete, auditable decision rules. The upper table specifies the numeric boundaries at each interim: at N = 20, the trial declares Go if efficacy events >= 6 and toxicity events <= 5; it stops for futility if efficacy events <= 2; and it stops for toxicity if toxicity events >= 8, regardless of efficacy. At N = 35 and N = 50, the boundaries scale accordingly. This tabulation eliminates ambiguity — at any interim sample size, the team knows exactly which (n_eff, n_tox) pairs trigger which decision. The lower heatmap visualises the same logic in two-dimensional space for a representative N = 40. The horizontal axis spans efficacy events, the vertical axis spans toxicity events; each cell is colour-coded by decision outcome — green (lower-right, GO: sufficient efficacy and controlled toxicity), blue (left, STOP: futility — efficacy too low regardless of toxicity), red (upper, STOP: excess toxicity — toxicity too high regardless of efficacy), grey (middle, Continue: insufficient evidence, continue enrollment). The heatmap reveals two core structural features of the Eff/Tox framework. First, the red toxicity-stop band spans the full width of the plot — once toxicity exceeds the boundary, the trial stops regardless of how many efficacy events are observed. This is the visual embodiment of toxicity-override priority: patient safety cannot be 'offset' by high efficacy. Second, the green GO region is confined to the lower-right quadrant — the decision imposes simultaneous constraints on both efficacy and toxicity; dominance in a single dimension is never sufficient for a Go determination. The large grey area between them represents the appropriate response to insufficient evidence: continue enrollment, rather than making a premature decision under uncertainty. This table-plus-heatmap combination, which pre-partitions (n_eff, n_tox) space into four mutually exclusive decision regions, essentially provides the trial with an automated navigation rule: no re-discussion or re-interpretation at each interim — the boundaries were pre-specified and visualised before the first patient was enrolled.

Figure 14. Eff/Tox operating characteristic analysis: four key scenario summary cards (H0: Low Eff/Safe, H1: Target Eff/Safe, H0+Toxic, H1+Toxic) with P(Go), P(futility stop), P(toxicity stop), and

mean sample size; global Go probability heatmap across (true ORR%, true G3+ AE%) with decision threshold reference lines.

Figure 14 presents the OC simulation results that validate the design's operating characteristics before any patient data are collected. The four summary cards at top span the clinically relevant scenario space. Card 1 — H0: Low Eff / Safe (true ORR = 10%, true G3+ AE = 20%): P(Go) = 0.9%, confirming strict Type I error control — an ineffective drug is almost never advanced. The majority of trials (88.4%) stop for futility. Card 2 — H1: Target Eff / Safe (true ORR = 35%, true G3+ AE = 20%): P(Go) = 89.6% with mean N = 27 — the design achieves approximately 90% power, and the mean sample size is well below the maximum of 50 because many simulated trials stop early with sufficient evidence. Card 3 — H0 + Toxic (true ORR = 10%, true G3+ AE = 40%): P(Toxicity stop) = 63.6% — the design reliably detects excessive toxicity, protecting patients from a drug that is both ineffective and unsafe. Card 4 — H1 + Toxic (true ORR = 35%, true G3+ AE = 40%): this is the critical scenario where efficacy and toxicity conflict. P(Go) drops to 14.3% while P(Toxicity stop) rises to 79.3% — even though the drug is genuinely efficacious, the toxicity signal dominates, and the design correctly recommends stopping in the majority of trials. The lower heatmap provides a global view: P(Go) across the full grid of (true ORR%, true G3+ AE%), with dashed reference lines at λ_E = 20% and λ_T = 30%. The green region (high Go probability) is concentrated in the lower-right — high efficacy, low toxicity — confirming that Go decisions require both conditions simultaneously. The red-to-green transition across the middle diagonal demonstrates the design's sensitivity: it rewards genuine efficacy and penalises toxicity, never conflating the two.

Eff/Tox sequential monitoring and BOIN12 serve complementary roles in the dose-optimisation workflow. BOIN12 answers 'among several doses, which one maximises the efficacy-toxicity utility trade-off?' — a dose-selection question for Phase I. Eff/Tox answers 'at this specific dose, does the accumulating efficacy-toxicity evidence support continuation, stopping for futility, or stopping for toxicity?' — a sequential monitoring question for Phase II. A development programme may use both: BOIN12 in Phase I to identify the OBD, then Eff/Tox sequential monitoring in Phase II to confirm that the OBD maintains an acceptable efficacy-toxicity profile as more patients are treated. Compared with Simon Two-Stage — which asks only 'is the ORR high enough?' — Eff/Tox asks 'is the ORR high enough AND is the toxicity low enough?' — a richer question that directly addresses the FDA's emphasis on benefit-risk assessment in Project Optimus. The module is implemented in R/efftox_ph2.R (decision boundary heatmap, OC simulation, live posterior updating), with the Shiny interface in R/server/server_efftox.R.

DECISION SUPPORT

The Decision Support domain is the hub where cross-domain evidence converges into a single, auditable recommendation. Following the BRAT (Benefit-Risk Action Team) structured assessment framework, the platform synthesises evidence from all upstream analysis modules into a five-dimensional continuous Go/No-Go scorecard with SMAA weight sensitivity analysis (§8.2). Evidence auto-propagates from upstream modules: the ORR posterior that drives the scorecard is the identical posterior from the Bayesian Go/No-Go module. No separate numbers need to be entered, eliminating the transcription errors that arise when statistical outputs are manually copied into presentation slides. The decision output maps to a four-tier recommendation scale and is persisted to PostgreSQL with a complete audit trail supporting 21 CFR Part 11 compliance.

8.1 COMPETITIVE BENCHMARKING AND MAP PRIOR

The Competitive Benchmark module (R/benchmark_analysis.R) answers a core question: where does the current trial's ORR and safety profile stand in the competitive landscape? The platform maintains a competitor database (competitor_benchmark.csv) in which each record captures the drug name, target, indication, trial phase, sample size, ORR (with 95% CI), mPFS/mOS, Grade >= 3 AE rate, and data source. Users can filter dynamically by target class, line of therapy, and trial phase. The interactive benchmark dashboard (Figure 15) contains seven tabs: Database (competitor registry), Efficacy (ORR/CRR forest plot and TTE bar chart), Safety (four safety indicators as horizontal bar charts and heatmaps), Positioning (2D bubble chart, RBI vs G3+ rate), MAP Prior (historical borrowing and posterior comparison), NMA (Network Meta-Analysis), and Intelligence (competitive signal detection and AI clustering). Three functions power the analysis. compute_map_prior() transforms historical data into a Beta prior via weighted pooling with a robust mixture component; compute_posterior_with_map()

performs conjugate updating with current trial data and detects prior-data conflict (z-score and predictive tail probability); compute_map_sensitivity() displays P(Go) robustness across a grid of delta values. All analyses can be populated with one click via the Auto-fill from Trial Data button, which pulls results from the upstream Bayesian module.

The competitive benchmark module feeds the Go/No-Go scorecard (§8.2) through two pathways. First, the Differentiation dimension: the user identifies the best competitor ORR in the benchmark dashboard, enters it into the scorecard's competitor input field, and score_differentiation() computes the posterior probability P(ORR > best competitor). Second, the MAP prior pathway: compute_posterior_with_map() transforms historical data into a Beta prior, performs conjugate updating with current trial data, participates directly in P(Go) recalculation, and detects prior-data conflict. Together, these two pathways replace the anecdotal governance-meeting exchange ('I recall the competitor's ORR was about 30%') with auditable percentile rankings and quantitative historical-borrowing evidence.

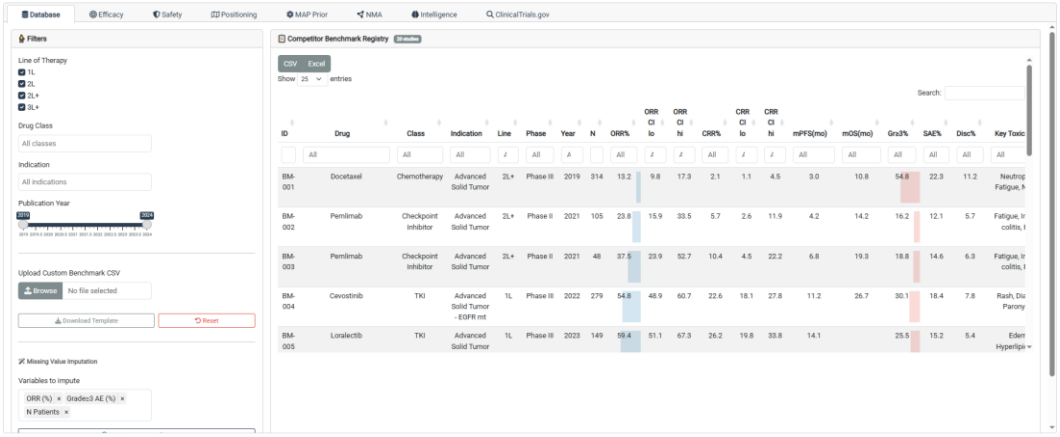

Figure 15. Competitive Benchmark Dashboard — Database tab: left filter panel — line of therapy (checkboxes), drug class and indication (multi-select dropdowns), publication year (slider), and upload/download/reset buttons, with missing-data imputation (RF/missForest) variable selection and execute button below; right — Competitor Benchmark Registry data table (DT interactive sort/filter/full-screen), columns include drug name, target, indication, trial phase, sample size, ORR with 95% CI, mPFS, mOS, Grade >= 3 AE rate, and data source.

8.2 FIVE-DIMENSIONAL GO/NO-GO CONTINUOUS SCORECARD WITH SMAA

The five-dimensional Go/No-Go scorecard (R/go_nogo_decision.R) extends the binary Bayesian decision into five continuously scored dimensions, each yielding a P(Go) between 0 and 1 and a weighted contribution to the overall score: (1) Efficacy: P(ORR > TPP target), based on a Beta posterior; (2) Safety: 1 - P(Grade >= 3 TEAE rate > benchmark), based on a Beta posterior, toggleable to SAE or DLT; (3) Differentiation: P(ORR > best competitor), based on Monte Carlo sampling from two independent Beta posteriors; (4) Feasibility: a product score combining enrollment velocity, regulatory feasibility, and trial readiness; (5) Commercial: a product score combining market size, remaining patent life, and competitive intensity. The default weights are 30%, 25%, 20%, 15%, and 10% respectively, and may be adjusted in the trial configuration. The weighted total from compute_gonogo_decision() maps to a four-tier decision: >= 0.60 Strong Go, 0.40-0.60 Conditional Go, 0.25-0.40 Need More Data, < 0.25 No-Go. Weight uncertainty is quantified via SMAA (gonogo_smaa()): 10,000 random weight vectors are sampled from a Dirichlet distribution, and for each dimension an acceptability index is computed, defined as the proportion of scenarios in which that dimension is the decisive factor.

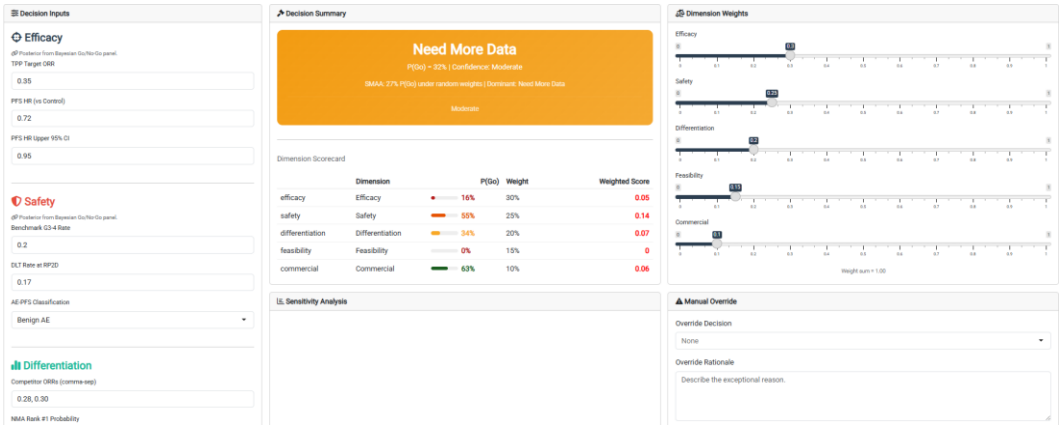

Figure 16. Five-Dimensional Go/No-Go Scorecard with SMAA dashboard: left — five-dimension scorecard (Efficacy, Safety, Differentiation, Feasibility, Commercial), each displaying P(Go) and confidence rating; upper-right — weighted total and SMAA acceptability indices; lower-right — four-tier decision mapping (current level highlighted).

Figure 16 illustrates a classic Need More Data case: the weighted total P(Go) = 32%, but the SMAA acceptability index is only 27%, lower than the nominal total and indicating that the decision is not robust under weight perturbations. The dimensional decomposition reveals the root cause. Efficacy P(Go) is only 16% (weight 30%), the primary drag on the total score. Safety at 55% (25%) and Commercial at 63% (10%) are acceptable. Feasibility, however, has a P(Go) of 0% (weight 15%), representing the most dangerous signal: severe obstacles exist in enrollment or the regulatory pathway. This decomposition transforms the decision discussion from a binary 'Go or No-Go' debate into a structured diagnosis of which dimensions contain evidence gaps that need priority resolution.

The scorecard's governance value is its pre-specification architecture: the weights, SMAA parameters, and decision mapping are all locked before data unblinding. The SMAA acceptability index provides the critical sensitivity check. If no single dimension exceeds 50% acceptability, the decision is genuinely balanced across dimensions and warrants deeper discussion rather than a formulaic recommendation. The module is implemented in R/go_nogo_decision.R, with the Shiny interface in R/server/server_gonogo.R.

8.3 AI-AUGMENTED DECISION MEETING REPORT: FROM ANALYSIS FRAGMENTS TO A REGULATORY-GRADE DOCUMENT PACKAGE

The analysis modules described in the preceding sections (Safety Analysis, §4; Efficacy Assessment, §5; PK/PD, §6; Dose Optimization, §7; Go/No-Go Scorecard, §8.2; and Competitive Benchmark, §8.1) each produce independent numbers and figures. While individually complete, these outputs are scattered across multiple tabs during decision meetings, forcing DSMB members, regulatory reviewers, and governance committees to flip between interfaces and manually assemble the complete evidence picture. The Decision Meeting Summary Report module (R/report_module.R, R/server/server_report.R) fills this gap. It captures the latest posterior probabilities, safety counts, dose decisions, and ML risk stratifications in real time from the reactive data streams of all upstream modules and compiles them into a structured, downloadable, standalone HTML file. The report is formatted to align with the regulatory standards of ICH E3 (Clinical Study Report) and ICH E9(R1) (Estimand framework), making it directly usable for DSMB/DMC meetings, internal Go/No-Go reviews, dose-escalation reviews, and regulatory briefings.

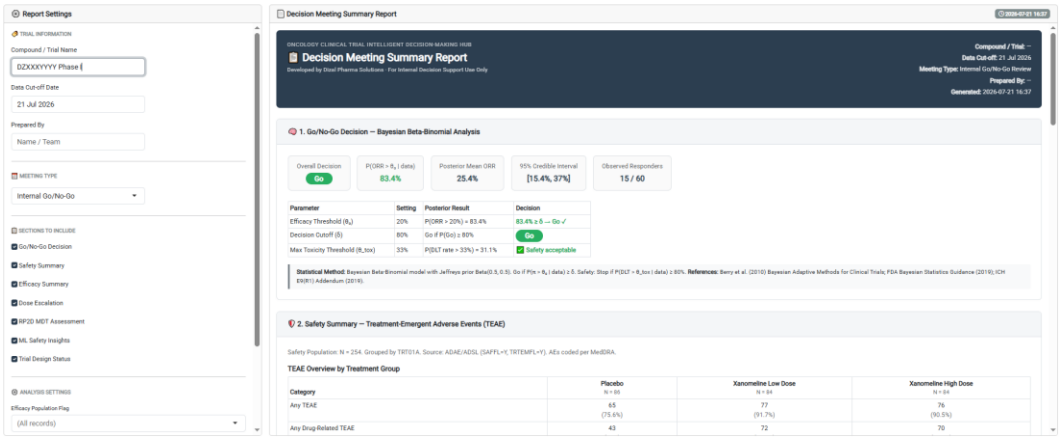

Figure 17. AI-Augmented Decision Meeting Report: left — report settings panel (trial information, meeting type, section checkboxes); right — rendered structured HTML report containing Go/No-Go decision, safety summary, efficacy summary, dose escalation, RP2D MDT assessment, ML safety insights, integrated recommendation, and trial design status.

Report generation is user-triggered. When the user clicks 'Generate Report', the report_snapshot() event captures a one-time snapshot of the current state from all upstream modules and passes it to render_decision_report(), which renders the following 9 standard sections:

1. Report Header — trial name, data cutoff date, meeting type (DSMB / Go-No-Go / Dose Review / Regulatory Briefing / Investigator Meeting), and generation timestamp;
2. Go/No-Go Decision — Bayesian Beta-Binomial analysis output, including P(ORR > θ_0 | data), posterior mean ORR, 95% credible interval, observed response count, and a binary Go/No-Go recommendation based on a pre-specified δ cutoff;
3. Safety Summary — TEAE summary table stratified by treatment arm (Any TEAE, Drug-related TEAE, Grade >= 3, Drug-related Grade >= 3, SAE, AE leading to dose modification, AE leading to discontinuation, Fatal AE), reporting n/N(%) per arm; plus DLT summary table stratified by dose;
4. Efficacy Summary — frequency distribution table of ORR and BOR (Best Overall Response) by assessor and population flag, with Clopper-Pearson exact 95% confidence intervals;
5. Dose Escalation and RP2D MDT — dose-escalation design type (BOIN/mTPI/3+3/CRM), target DLT rate, DLT summary by dose, Bayesian CRM+EWOC MTD posterior estimate with MCMC convergence diagnostics; and Multi-Criteria Decision Analysis (MCDA) output with pass/fail status for each constraint and integrated recommendation narrative;
6. ML Safety Insights — Logistic regression patient risk score distribution (based on 5 AE features: n_TEAE, max grade, drug-related events, SAE, dose modification; patient count and percentage by Low/Medium/High risk tier), K-means cluster labels and sizes with dominant features, CUSUM monitor current status and most recent alerts (if any);
7. Integrated Decision Recommendation — domain scoring table (with weights, percentage scores, and Go/No-Go status), weighted evidence maturity bar chart, list of key integrated findings;
8. Trial Design Status — Simon Two-Stage design parameters (n1/n2, r1/r, EN(p_0), PET(p_0)) and dose-escalation design type and target DLT rate;
9. Disclaimer — internal decision-support use statement, with citations of applicable regulatory guidance (FDA Bayesian Statistics Guidance 2019, ICH E9(R1) 2019, ICH E3, ICH E6(R3) GCP 2023).

Each section is displayed only when the user has checked that section and the corresponding upstream analysis has been completed. Analyses not yet run are annotated as 'Not yet run' rather than silently omitted, giving readers full transparency about the report's completeness.

The report module's design follows three architectural principles. First, separation of concerns: report_snapshot() is responsible only for data capture and format standardisation, while render_decision_report() handles only HTML rendering. The two functions communicate through a pure data list (snap), which means the rendering logic can be independently tested without relying on the Shiny

reactive context and the same snapshot can be reused by different renderers (HTML, PDF, Word). Second, regulatory alignment: each section's footer annotates the corresponding ICH guideline clause (e.g., the safety section notes 'Per ICH E3 section 12.3.2'; the efficacy section notes 'Per ICH E3 section 11.4.1'), allowing reviewers to map report content directly to regulatory requirements. Third, auditability: the report header contains a minute-precision timestamp, trial name, and data cutoff date. The generated HTML file is a self-contained document (inline CSS, Bootstrap via CDN) that can be preserved as part of the decision audit trail. The report supports one-click download as a standalone HTML file; the filename automatically includes the trial name and date, making it suitable for direct attachment to meeting materials, email distribution, or archiving in the electronic Trial Master File (eTMF). The module is implemented in R/report_module.R (render_decision_report) and R/server/server_report.R (report_snapshot, download handler), with the Shiny interface located in the 'Summary Report' tab under the 'More' navigation menu.

CASE STUDY: END-TO-END DECISION WORKFLOW

The preceding chapters have laid out the complete methodological framework, from safety analysis through full lifecycle management. But methods reveal their true value only in real cases. This chapter weaves these methods into an end-to-end decision chain through a fictional yet developmentally realistic case: the Phase I-II clinical development of X-247, a novel oral tyrosine kinase inhibitor (TKI) targeting EGFR exon 20 insertion mutations in advanced non-small cell lung cancer (NSCLC) after failure of platinum-based chemotherapy. This mutation subtype currently has only two approved drugs (ORR 28-30%), representing both a clear unmet clinical need and a differentiation opportunity. Before Phase I began, the project team established a preliminary Target Product Profile (TPP): target ORR >= 35% (vs. competitors at 28-30%), median PFS >= 8 months, Grade 3-4 treatment-related AE rate < 20%, capsule formulation, and once-daily oral administration.

9.1 PHASE I: BOIN DOSE ESCALATION AND MTD DETERMINATION

Phase I employed a BOIN design (target DLT rate 0.25) with five pre-specified dose levels (40, 80, 120, 160, 200 mg QD) and a planned enrollment of up to 30 patients. The dose escalation proceeded as follows: 40 mg (0/3 DLT) escalated to 80 mg (0/3 DLT), which escalated to 120 mg (1/6 DLT, rate 16.7%, within the BOIN retention interval), prompting the design to retain at this dose. The next cohort at 160 mg recorded 2/6 DLTs (33.3%), triggering de-escalation back to 120 mg. BOIN decision rules thus identified 120 mg as the MTD, with 160 mg exceeding the safety boundary.

While the MTD was being determined, the system ran three supplementary analyses to inform the RP2D decision. For safety, the most common AEs at 120 mg were Grade 1-2 rash (67%) and diarrhoea (50%); EAIR analysis showed no significant excess over historical benchmarks for the same TKI class (system module: eair_analysis.R). For exposure-response, AUC increased linearly with dose (R squared = 0.91), and PK quartile analysis showed a response rate of 42% in Q3-Q4 versus 11% in Q1-Q2 (system module: exposure_response_ml.R). For competitive intelligence, two contemporaneous competitors had Phase II ORRs of 28% and 30% (system module: benchmark_analysis.R).

Integrating all four evidence streams, the decision committee approved 120 mg as the RP2D. This was not a single-dimension judgment. BOIN answered the question 'where is the MTD?' by identifying the dose whose toxicity probability is closest to the target DLT rate of 25%. But BOIN alone was insufficient to constitute the RP2D. Three additional lines of evidence, the E-R analysis, the EAIR benchmark comparison, and the competitive landscape assessment, together advanced the conclusion from '120 mg is the MTD' to '120 mg has a preliminary efficacy signal with manageable safety.'

9.2 PHASE II: RANDOMISED DOSE OPTIMIZATION — FROM MTD TO OBD

The 120 mg dose identified by Phase I BOIN was the MTD, defined as the dose whose toxicity probability was closest to the target DLT rate of 0.25. But the MTD logic has a fundamental limitation: it considers only toxicity, implicitly assuming that higher dose equals more target inhibition and therefore better efficacy. In the targeted therapy era, this assumption has been repeatedly disproved: higher doses beyond target saturation increase toxicity without adding efficacy. Before locking 120 mg as the dose for

confirmatory trials, the team therefore chose not to default to the MTD. Instead, they conducted a small randomised comparison to answer two questions. Does 120 mg genuinely offer superior efficacy over the next lower dose of 80 mg? And if efficacy is comparable, should the lower, safer dose be preferred?

Based on Phase I data, both 80 mg and 120 mg were within the safety range (DLT rates of 0% and 16.7%, respectively), each below the 25% target. E-R analysis confirmed that AUC increased linearly with dose ($R^2 = 0.91$), but the efficacy difference between 80 mg and 120 mg had not been directly compared. The team therefore designed a small randomised dose comparison: 1:1 randomisation, open-label, two parallel arms (80 mg QD vs. 120 mg QD), 20 patients per arm (40 total), with a primary endpoint of ORR by independent radiology review (RECIST 1.1) and secondary endpoints including Grade 3+ treatment-related AE rate, DLT rate, PK exposure (AUC_{0-24}), and duration of response (DOR). Three analysis modules were employed: two-arm Bayesian superiority monitoring (two_arm_bayes.R), Miettinen-Nurminen stratified comparison (miettinen_nurminen.R), and E-R plateau analysis (exposure_response_ml.R). The pre-specified decision criteria were as follows: if $P(ORR_{120} > ORR_{80} + 0.05 \mid data) >= 0.80$, confirm 120 mg as the preferred dose; if the reverse held at the same threshold, switch to 80 mg; otherwise, select based on a composite safety and E-R score.

All 40 patients completed at least one tumour assessment cycle (data cutoff at Week 12), with balanced baseline characteristics between arms. On efficacy, 120 mg showed a numerically higher ORR than 80 mg (40% vs. 35%, an absolute difference of +5%), but the difference was small and the 95% CI was wide — the characteristic signature of a small-sample randomised comparison: directional indication rather than confirmatory evidence. On safety, 120 mg was consistently worse than 80 mg across all AE dimensions: the Grade 3+ TRAE rate rose from 10% to 15%, Grade 1-2 rash from 35% to 55%, and diarrhoea from 25% to 45%. This pattern suggested that in the 80-120 mg interval, the dose-toxicity relationship was still rising while the dose-efficacy relationship was approaching a plateau. The main results are summarised in Table 9-1.

Endpoint	80 mg (n=20)	120 mg (n=20)	Difference (120-80 mg)
ORR (CR+PR)	7/20 (35%)	8/20 (40%)	+5% (95% CI: -23% to +31%)
Median DOR (months)	7.8 (95% CI: 5.1-NE)	9.2 (95% CI: 6.4-NE)	—
G3+ TRAE rate	2/20 (10%)	3/20 (15%)	+5% (95% CI: -17% to +27%)
DLT rate	0/20 (0%)	2/20 (10%)	+10%
Most common AE (G1-2 rash)	7/20 (35%)	11/20 (55%)	+20%
Most common AE (G1-2 diarrhoea)	5/20 (25%)	9/20 (45%)	+20%
AUC_{0-24} (ng·h/mL)	2850 ± 620	4120 ± 890	—

The two-arm data were entered into the Two-Arm Bayesian Superiority module (system menu: Trial Design -> Two-Arm Superiority), with the minimum clinically significant difference (MDD) set at 5 percentage points, a Beta(0.5, 0.5) Jeffreys prior, and 20,000 Monte Carlo samples. The system output the following: 120 mg arm posterior mean ORR 40.5% (8/20), 80 mg arm posterior mean ORR 35.7% (7/20). $P(\Delta > 0.05 \mid data) = 0.49$ — the posterior probability that 120 mg exceeds 80 mg in ORR by at least 5 percentage points was less than half, well below the pre-specified 80% threshold for confirming superiority. Setting MDD = 0 yielded $P(\Delta > 0 \mid data) = 0.63$ — a simple numerical comparison gave a 63% posterior probability that 120 mg has a higher ORR than 80 mg, a weak directional signal. The Miettinen-Nurminen method provided a frequentist corroboration: ORR difference +5% (95% CI: -24.7% to +33.8%), p = 0.747 (two-sided). Taken together, these 40-patient data were insufficient to adjudicate decisively between the two doses — the direction of effect pointed weakly toward 120 mg, but the magnitude fell far short of confirmatory.

Faced with a result of 'efficacy difference not significant, safety directionally worse,' the decision committee chose to maintain 120 mg — but this was not an inertial 'default to MTD.' It rested on four

systematic trade-offs. First, the directional ORR difference of +5%, while not statistically significant, is a meaningfully sized signal in a sample of n = 20 per arm — if the true difference were zero, the prior probability of observing an absolute difference of at least +5% would be approximately 35-40%. Second, the magnitude of the safety difference did not yet constitute a clinical concern: the Grade 3+ TRAE rate increase from 10% to 15% remained within the TPP-specified '< 20%' ceiling, and both Grade 1-2 rash and diarrhoea are manageable with supportive care. Third, the key E-R finding was that the efficacy E-R curve had not yet plateaued in the 80-120 mg interval — the AUC increase from 2,850 to 4,120 ng.h/mL (+45%) was accompanied by an ORR rise from 35% to 40%, a pattern different from the zongertinib case, in which the 120 vs. 240 mg interval showed higher ORR but shorter response. X-247 at 120 mg thus still held potential for further efficacy gain. Fourth, and most critically, the value of this randomised comparison lay not in proving that 120 mg is significantly better than 80 mg, but in (a) ruling out the definitive conclusion that 80 mg is non-inferior to 120 mg (P = 0.62, not P = 0.10), and (b) providing a randomised data foundation for the Phase III dose justification. Final decision: 120 mg was confirmed as the dose for subsequent confirmatory development, while 80 mg was retained as a potential dose-reduction alternative.

9.3 BENEFIT-RISK ASSESSMENT AND GO/NO-GO DECISION

The results of the randomised dose comparison — 120 mg arm ORR 40% (8/20), Grade 3+ TRAE rate 15% (3/20) — were entered directly into the benefit-risk assessment module (server_benefit_risk.R + server_go_nogo.R). The BRAT evidence table summarised the evidence across three key dimensions: Efficacy (ORR 40%, vs. TPP 35% and vs. competitors 28-30%), Safety (Grade 3+ AE rate 15%, vs. TPP < 20% and vs. competitor 22%), and Differentiation (delta-ORR = +5%). SMAA analysis showed that 120 mg X-247 ranked first in 82% of weight samples, compared with Competitor A at 18%. The sensitivity grid confirmed that this conclusion remained robust within an ORR range of +/- 5 percentage points — regardless of whether the decision-maker weighted efficacy or safety more heavily, 120 mg consistently ranked first.

The five-dimensional Go/No-Go scorecard returned the following: Efficacy 4.2/5 (ORR close to TPP and ranked first in the competitive landscape), Safety 3.8/5 (Grade 3+ AE rate met the target but diarrhoea required proactive management), Differentiation 4.0/5, Feasibility 3.5/5 (oral capsule, once-daily), Commercial 3.5/5. Weighted total: 3.9/5 — triggering 'Conditional Go,' with recommendations to initiate Phase III, add prophylactic anti-diarrhoeal strategies, and pre-specify AE management guidelines in the Phase III protocol. Unlike the traditional approach of checking whether ORR exceeds a single pre-specified threshold, the Go/No-Go determination here integrated five dimensions and verified robustness to weight variation through SMAA. The decision committee approved unanimously, and the decision was recorded in the system audit log together with the complete traceability chain: Bayesian posterior probabilities, SMAA acceptability indices, and committee voting records.

IMPLEMENTATION, VALIDATION, AND DISCUSSION

10.1 STATISTICAL METHOD VALIDATION: ANALYTICAL VERIFICATION, SIMULATION, AND CROSS-VALIDATION

The platform's validation strategy rests on a single principle: every statistical method must pass three tiers of verification — analytical agreement with closed-form formulae, numerical cross-validation against published methods, and frequentist calibration through Operating Characteristic (OC) simulation. The validation infrastructure comprises 13 testthat test files, 310+ independent test cases, 4 automated validation scripts, and 7 GAMP 5 validation documents, all located under tests/testthat/ and validation/.

Analytical verification forms the first line of defence. For all Bayesian modules using conjugate priors, test cases directly call R probability functions (pbeta(), pbinom(), pgamma(), qnorm(), qchisq()) to compute expected values and compare them against function outputs at tolerances ranging from 1e-4 to 1e-12. For example, the Go/No-Go tests (test-bayesian_stats.R, BAS-TC-01 through TC-20) verify that compute_posterior_go(7, 20, 0.2, 0.8) yields P(Go) = 0.9993 ± 0.0001; the safety posterior tests verify the closed-form result for P(toxicity rate > 0.33) under a Jeffreys prior Beta(0.5, 0.5) with 4 observed DLTs out of 12 patients. For Simon Two-Stage designs (test-simon_two_stage.R, SIM-TC-01 through TC-36),

the rejection region of each design (n1, r1, n2, r) is verified exactly via pbinom() and compared directly against the published results in Simon (1989) Table 3.

Cross-validation targets modules that implement published algorithms. The BOIN MTD selector (test-boin_mtd_validation.R) directly compares select_mtd_boin() against the CRAN BOIN package's select.mtd() under standard scenarios — the test file preserves CRAN package output per scenario in comments for audit traceability (e.g., '# BOIN::select.mtd(target=0.25, npts=c(6,6,6), ntox=c(0,0,0))$MTD == 3'). The mTPI decision table (test-dose_escalation.R) verifies the isotonic regression upper and lower bounds row-by-row against the UPM critical values in Ji et al. (2010) Table 1. The EffTox boundary table (test-efftox_ph2_validation.R) independently verifies the efficacy-toxicity trade-off boundary roots of Thall and Cook (2004) via uniroot(). The BOIN-Combo two-dimensional lambda boundary values (test-combo_boin_validation.R) are compared against the analytical expressions of Lin and Yin (2017).

OC simulation is the third tier — not validating code correctness, but confirming that the statistical method's behaviour matches its claimed design parameters. Every module involving dose selection or Go/No-Go decisions includes OC simulation functions: simulate_boin() runs 2,000 virtual trials under multiple true toxicity scenarios, computing the correct MTD selection probability and the average proportion of patients allocated to excessively toxic doses; efftox_oc_grid() runs parallel EffTox trial simulations across an 11 x 11 (true ORR%, true Grade 3+ AE%) grid, verifying the monotonicity of the P(Go) heatmap (P(Go) should increase monotonically as true efficacy increases); btte_oc_grid() performs the same verification for Bayesian TTE designs; ph23_oc_sim() computes conditional error rates, Fisher combination test chi-squared(4) distribution consistency, and Winner's Curse correction effectiveness for seamless Phase II/III designs. All simulations use fixed random seeds for reproducibility.

10.2 GAMP 5 VALIDATION FRAMEWORK AND AUTOMATED OQ/PQ

The platform was constructed under a three-stage (3Q) validation lifecycle in accordance with GAMP 5 Category 4 (Configurable Software), comprising 7 formal validation documents (all bilingual Chinese-English, in .md + .docx format). The Validation Plan (VP-OCTDM-001) defines system scope, role responsibilities, acceptance criteria, and deviation management procedures. The User Requirements Specification (URS) maps the five clinical function domain requirements to testable functional specifications. The Configuration Specification (CS) documents hardware specifications, operating system configuration, R runtime environment, and all R package versions. Installation Qualification (IQ, verify_os.sh + verify_r_packages.R) verifies compliance at two levels: the operating system level (RHEL/Ubuntu version, R >= 4.5, BLAS library, SELinux/AppArmor status, locale settings, NTP synchronisation, Shiny service status, CPU and memory baseline) and the R package level (inventory completeness, minimum version requirements, functional smoke tests). Operational Qualification (OQ, oq_test_suite.R) contains 6 suites and 28 test cases (22 automated + 6 manual browser interactions), covering the R environment (3 TCs), CDISC data layer (4 TCs), Safety modules (3 TCs), Bayesian statistics (6 TCs), PK/ER modules (3 TCs), and report generation (3 TCs), with the acceptance criterion that all mandatory cases must pass and the overall pass rate must be at least 95%. Performance Qualification (PQ, pq_benchmark.R) contains 7 performance benchmarks with quantitative thresholds: data loading <= 30 seconds, AE summary table <= 10 seconds (3,000 ADAE rows), Bayesian posterior computation <= 5 seconds, Word report generation <= 60 seconds, response latency under 5 concurrent sessions < 2x baseline, memory usage <= 80% RAM, BOIN 2,000-simulation OC <= 120 seconds. The Validation Summary Report (VSR) aggregates results, deviations, and the final pass/fail determination across all 3Q stages. The complete set of validation documents is designed to support Computerised System Validation (CSV) evidence in regulatory submissions — although the current version is marked 'Draft,' the framework structure and automated scripts are directly executable.

10.3 DECISION AUDIT TRAIL AND DATA PERSISTENCE

A clinical trial decision-support platform operating in a regulatory environment must satisfy two non-negotiable requirements: every decision must be traceable to the exact data, parameters, and user at the time the decision was made; and analysis results must persist across sessions to support trend analysis and audit review across data cutoff dates. The platform's audit trail layer (R/db_connection.R, 1,032 lines) is implemented via PostgreSQL, with a graceful degradation pathway — when the environment variable

PG_HOST is not set, the system automatically falls back to CSV/in-memory mode, ensuring operability in local development environments without database infrastructure.

The database schema consists of five core tables that together form a complete decision evidence chain:

1. decision_audit_log — an immutable decision log recording a snapshot of every Go/No-Go decision. Fields include: trial_id (supporting multi-trial deployment), logged_at (TIMESTAMPTZ, microsecond-precision timestamp), decision (Go/No-Go determination), responders (observed responder count / n total), p_go_pct (P(Go) percentage), p_dlt_tox_pct (P(DLT > toxicity threshold) percentage), stage (trial phase), notes (free-text annotation), user_role (role triggering the decision — Medical/Pharmacology/Statistics), analysis_params (JSONB — the complete parameter set triggering this decision, including threshold θ_0, decision cutoff δ, and toxicity threshold), data_md5 (MD5 fingerprint of the input CDISC data, ensuring input data traceability), and override_flag (boolean — marking whether this decision was manually overridden). The INSERT-only design (no UPDATE operations) ensures log immutability.

2. evidence_score_history — evidence score history, automatically written by the decision synthesis engine (§8.2) at each computation. Records overall_score (composite score 0-100), eff_status/safe_status (Go/No-Go status for the efficacy and safety dimensions), and domain_scores (JSONB — six-domain score snapshot). This table supports cross-time score trend analysis: auditors can query whether the safety score has been trending up or down across successive data cutoffs, detecting systematic drift in evidence quality.

3. analysis_results — a generic analysis result store serving as the unified persistence layer for all analysis modules. A UNIQUE INDEX ON (trial_id, analysis_key) ensures that only the most recent result is retained per analysis key (INSERT ... ON CONFLICT DO UPDATE). The ANALYSIS_KEYS constant defines standardised analysis key names (e.g., bayes_go_nogo, simon_design, dose_escalation_boin, efftox_tc), each key mapping to a parameter-and-result JSONB for that analysis type. This design enables report_snapshot() in server_report.R to recover previously computed results from the database: when an in-memory reactive value is NULL (new session), the system automatically falls back to db_read_analysis(key, trial_id), filling in missing analysis modules and ensuring that report generation is not lost due to session restart.

4. decision_scorecard — the complete five-dimensional scorecard audit, the highest-granularity record of decision auditing. Beyond total_score, decision, and confidence, it records the five dimensions' independent scores (eff_score through comm_score, on a 1-5 scale) and the weights used (w_eff through w_comm, to three-decimal precision), a sensitivity distribution summary based on a 56-cell sensitivity grid (JSONB of decision counts), raw input values (orr_observed, pfs_hr, ae_grade34_rate, dlt_rate, nma_rank_prob, regulatory_path), and override_decision / override_rationale (manual override action and justification). This structure ensures that any single scorecard result can be fully reproduced and independently audited after the fact.

5. cdisc_column_labels — a CDISC metadata preservation table. Because haven::zap_label() strips SAS variable labels at data import, this table captures and preserves column labels before writing to PostgreSQL, ensuring that the journey of data from SAS to PostgreSQL does not lose metadata.

The audit trail's Shiny interface is located in the 'Decision Audit Log' sub-tab of the Decision Center tab (§8.2). When the user clicks the 'Log Decision' button, the system captures the current Go/No-Go snapshot, parameters, and user role, and writes to the decision_audit_log table — simultaneously saving to CSV as backup. The bottom of the interface displays the cumulative decision history (in reverse chronological order), including a decision pattern mining panel: P(Go)% boxplots grouped by role (revealing decision threshold differences across Medical, Pharmacology, and Statistics roles), Go/No-Go stacked bar charts by trial phase (showing historical conservatism patterns), and a scatter plot of P(Go)% versus decision (with logistic regression estimating the empirical P50 — the P(Go)% tipping point where Go and No-Go are equally probable). When PostgreSQL is unavailable, audit logs and score histories are automatically saved to local CSV files (decision_audit_log.csv, evidence_score_history.csv). This dual-

track design — PostgreSQL as primary, CSV as backup — ensures that the basic requirements of an audit trail are met in any deployment scenario, while preserving the upgrade path to full 21 CFR Part 11 compliance.

CONCLUSION

This paper has presented an integrated, AI-augmented R Shiny platform for quantitative decision support in oncology clinical trials. The platform spans the complete decision chain from early safety signal detection (§4) to the Go/No-Go scorecard (§8.2), encompassing over 40 analysis modules across five clinical function domains. Three architectural decisions differentiate it from existing tools. First, all modules share a single reactive CDISC data layer, ensuring that the safety, efficacy, and decision-support domains operate on the exact same patient populations and analysis datasets — eliminating the common audit finding of inter-module data inconsistency. Second, a push-stream data architecture enables downstream modules, such as benefit-risk assessment and the AI signal engine, to read directly from upstream Bayesian posteriors rather than from manually copied values, preserving digital consistency from raw data to final recommendation. Third, the Decision Meeting Summary Report (§8.3) compiles all analysis outputs into a single structured HTML file aligned with ICH E3 and ICH E9(R1) directly usable for DSMB and DMC meetings — completing the last mile from data to decision to delivery.

The platform's validation framework (§10) is likewise a core contribution of this paper. Unlike analysis platforms that report only methodological descriptions, this paper demonstrates how each statistical module passes through three tiers of verification: analytical agreement with closed-form formulae, cross-validation against published methods (CRAN BOIN, Simon 1989 Table 3, Ji et al. 2010, Thall and Cook 2004), and frequentist calibration through operating characteristic simulation — totalling 310+ test cases distributed across 13 testthat test files. The GAMP 5 Category 4 validation lifecycle — comprising a complete 3Q (IQ/OQ/PQ) framework with 7 documents, 28 OQ test cases (with a >= 95% pass criterion), and 7 PQ performance benchmarks — provides a directly executable starting point for organisations wishing to adopt the platform for GxP environments. The immutable PostgreSQL audit trail layer (§10.3), supplemented by CSV backup, ensures that every Go/No-Go decision is fully reproducible: who made it, with what data, under what parameters, at what time.

As oncology drug development continues its trajectory toward smaller patient populations, more frequent interim decisions, and more complex benefit-risk trade-offs, the demand for tools that combine statistical rigour with operational agility will only grow. This paper demonstrates that such a platform is buildable — using open-source technologies, on a CDISC standards foundation, with rigorous validation that engages both the academic literature and the regulatory framework. The complete source code, STATISTICAL_METHODS.md (65KB Chinese statistical methodology documentation), and {testthat} validation suite are provided alongside this paper.

ACKNOWLEDGMENTS

The OncoTrialMind platform and the forthcoming book 《数据驱动的肿瘤临床试验决策》(Data-Driven Decision-Making in Oncology Clinical Trials) on which this paper is based were designed and developed by Huadan Li, which form the technical foundation and intellectual source of this work. The author expresses sincere gratitude for this foundational contribution. The author also thanks the clinical development and biostatistics teams who provided domain requirements and usability feedback during iterative development. The views expressed in this paper, and any errors that remain, are the sole responsibility of the author.

REFERENCES

Simon R (1989). Optimal two-stage designs for phase II clinical trials. Control Clin Trials 10: 1--10.

Koyama T, Chen H (2008). Proper inference from Simon's two-stage designs. Stat Med 27(16): 3145--3154.

Liu S, Yuan Y (2015). Bayesian optimal interval designs for phase I clinical trials. J R Stat Soc C 64: 507--523.

Lin R, Yin G (2017). Bayesian optimal interval design for dose finding with drug-combination trials. Stat Methods Med Res 26(5): 2154--2167.

Ji Y, Liu P, Li Y, Bekele BN (2010). A modified toxicity probability interval method for dose-finding trials. Clin Trials 7: 653--663.

Babb J, Rogatko A, Zacks S (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. Biometrika 85: 741--751.

Thall PF, Simon R, Estey EH (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. Stat Med 14: 357--379.

Zhou H, Murray TA, Pan H, Yuan Y (2019). BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. Clin Trials 16: 384--394.

Page ES (1954). Continuous inspection schemes. Biometrika 41: 100--115.

Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics 70(4): 1023--1032.

Jeffreys H (1946). An invariant form for the prior probability in estimation problems. Proc R Soc A 186: 453--461.

Thall PF, Cook JD (2004). Dose-finding based on efficacy-toxicity trade-offs using bivariate outcome data. Biometrics 60: 684--693.

— Further Reading: Regulatory Guidance and Policy Documents —

FDA (2019). Guidance for Industry: Bayesian Statistics in Medical Device Clinical Trials.

FDA (2026). Guidance for Industry: Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products.

ICH E9(R1) (2019). Addendum on Estimands and Sensitivity Analysis in Clinical Trials.

ICH E6(R3) (2023). Guideline for Good Clinical Practice.

EMA (2019). Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design.

China CDE (2020). Guidance for Industry on Adaptive Design in Clinical Trials.

China CDE (2023). Guidance for Industry on Patient-Focused Drug Development -- Clinical Trial Design, Implementation, and Benefit-Risk Assessment.

China CDE (2026). Guidance for Industry on Statistical Methodology for Patient-Centered Innovative Drug R&D (Draft for Comment).

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Zhiping Yan
Dizal Pharmaceutical Co., Ltd.
yanzhiping@dizalpharma.com
https://github.com/oncotrialmind